A Supervised Learning Method for Seismic Data Quality Control

Travis Addair
Lawrence Livermore National Laboratory

Background

At Lawrence Livermore National Laboratory (LLNL) the nuclear explosion monitoring group maintains a database of over 279,116,157 waveforms for 2,922,240 distinct events and 4,107 distinct stations. We had recently cross-correlated all possible pairs of waveforms for events within 50 km of one another and wanted to mine this dataset for statistics. However, an unknown number of the waveform segments contained non-seismic signal artifacts dominating the correlation, thus rendering the correlation value meaningless. These needed to be identified and removed before the project could proceed.

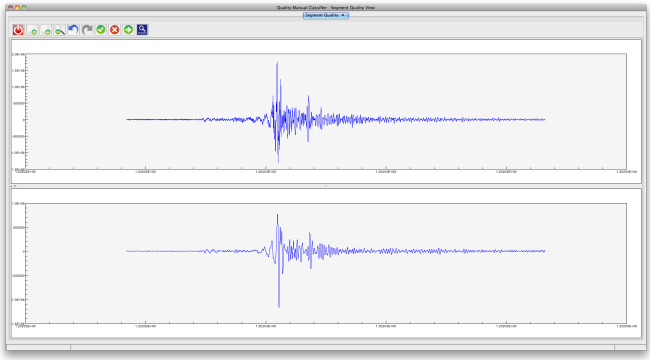
Manual Classification – Labeling

The initial cross correlation effort resulted in 15,710,298 correlations stored as rows in a database table. Each of these rows contained pointers to the waveform segments that were correlated, the band into which each waveform was filtered, and a series of relevant features for each of the two segments. Table 1 describes these features in greater detail.

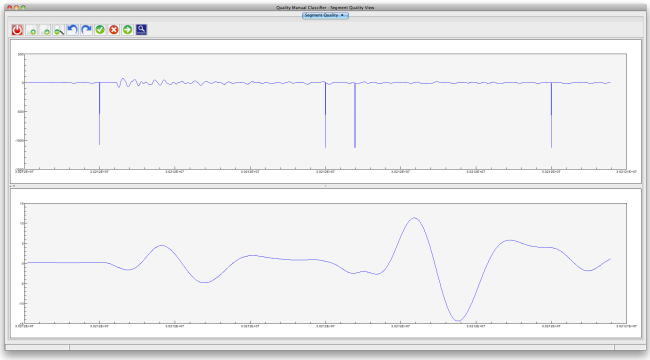
Table 1 Features used to classify segments

Abbreviation	Description
AVG_DELTA	Average of the distance (in degrees) of the catalog location of each event to the station
VARF	Time variance
VART	Frequency variance
T0	Time center
WINDOW_LENGTH	(Common) length of the correlation windows in seconds
LOW_CORNER	Low corner frequency of a bandpass filter (if applied)
HIGH_CORNER	High corner frequency of a bandpass filter (if applied)
SHIFT	Absolute shift in seconds of the second trace at the maximum
SIG_RMS	RMS computed in signal window
SW_SNR	Short-window signal-to-noise ratio
LW_SNR	Long-window signal-to-noise ratio
TBP	Time bandwidth product
POS_KURTOSIS	The sample kurtosis if positive, 0 otherwise (http://en.wikipedia.org/wiki/Kurtosis)
NEG_KURTOSIS	The sample kurtosis if negative, as a positive real number, 0 otherwise
EXTR_RAW	abs(median(signal) – mean(signal)) / range(signal)

In order to determine if this quality control problem would be well suited to machine learning techniques, we needed examples of both good and bad data from our correlation table. By visually examining the raw and filtered seismograms of a segment, a domain expert could discriminate between valid seismic data and invalid artifact-dominated data. To facilitate rapid classification of segments, a program was written to randomly sample from the over 15 million correlations in the table, and display the raw and filtered seismograms (Screens 1 and 2) for all the unlabeled waveform segments. When viewing each sample, the operator was given the choice to mark the segment as *good* or *bad*, after which the result was saved to a segment quality table for future use in automated classification.



Screen 2 Manual Classifier showing a good (valid) segment



Screen 1 Manual Classifier showing a bad (invalid) segment

Preprocessing

Unable to apply if-else rules to a satisfactory degree of accuracy, a supervised learning solution was explored. We selected the support vector machine (SVM) as our classification model of choice due its popularity as a binary classifier capable of efficiently operating in high dimensional feature spaces. In looking for third-party SVM implementations, our goals were to find an API that was open source and could easily integrate with our existing Java codebase, thereby facilitating interaction with the database and visualization of waveforms. The LIBSVM[1] package by Chih-Chung Chang and Chih-Jen Lin was selected for being the most prominent open source, off-the-shelf SVM with a Java implementation.

From the manual classification process, 23,648 unique waveform segment classifications of *good* (g) or *bad* (b) were stored in the segment quality table. After exporting the data and associated feature values to a flatfile, Python utility scripts provided by LIBSVM were used to validate the data, scale the features to a [-1, 1] range, and run a grid search to determine the optimal parameterization of the support vector machine using a Radial Basis Function (RBF) kernel. Literature provided by the authors of LIBSVM[2] suggests using the RBF kernel in most cases, particularly when the number of features is significantly smaller than the number of training examples. Having over 23,000 examples and only 15 features, this disparity clearly characterized our data set. Continued refinement of the grid search ultimately resulted in 5-fold cross validation accuracy of 95.35%.

Automatic Classification – Global Dataset

Using the parameters obtained from the grid search, an automatic classifier tailored specifically to our seismic data processing API was written as a wrapper for and extension to the LIBSVM Java library. In contrast to the original LIBSVM framework, this extension was able to process data by querying and updating the database directly. By fitting into our existing seismic processing API, the SVM was further able to integrate into other applications with minimal overhead.

To verify the correctness of our Java SVM implementation, holdout cross validation was performed on the labeled dataset, reserving 70% for training and 30% for testing. The accuracy for this test was similar to results from the cross-validation performed on the flatfile: around 95%. Given that the training data was skewed towards invalid examples 2:1, precision and recall were also calculated and determined to be approximately 92% and 91%, respectively, for segments classified as good using the same 70/30 split. From the standpoint of improving final correlation results, there were pros and cons to throwing more or less data out. Because of this ambiguity, accuracy was retained as the primary performance metric.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[2] http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

As a means of determining whether it would increase accuracy to gather more training data, we iteratively subtracted 500 samples from our 70% training data and measured our training and test accuracies on each of these subsets. As Figure 1 shows, we were clearly seeing diminishing returns as our training set increased, indicating convergence to a local maximum. Indeed, even after growing our labeled data set from 10,000 examples to 23,000, as mentioned above, we were unable to noticeably increase overall accuracy. We concluded that our misclassifications were the result of minor underfitting, and that additional features could be explored to improve performance.

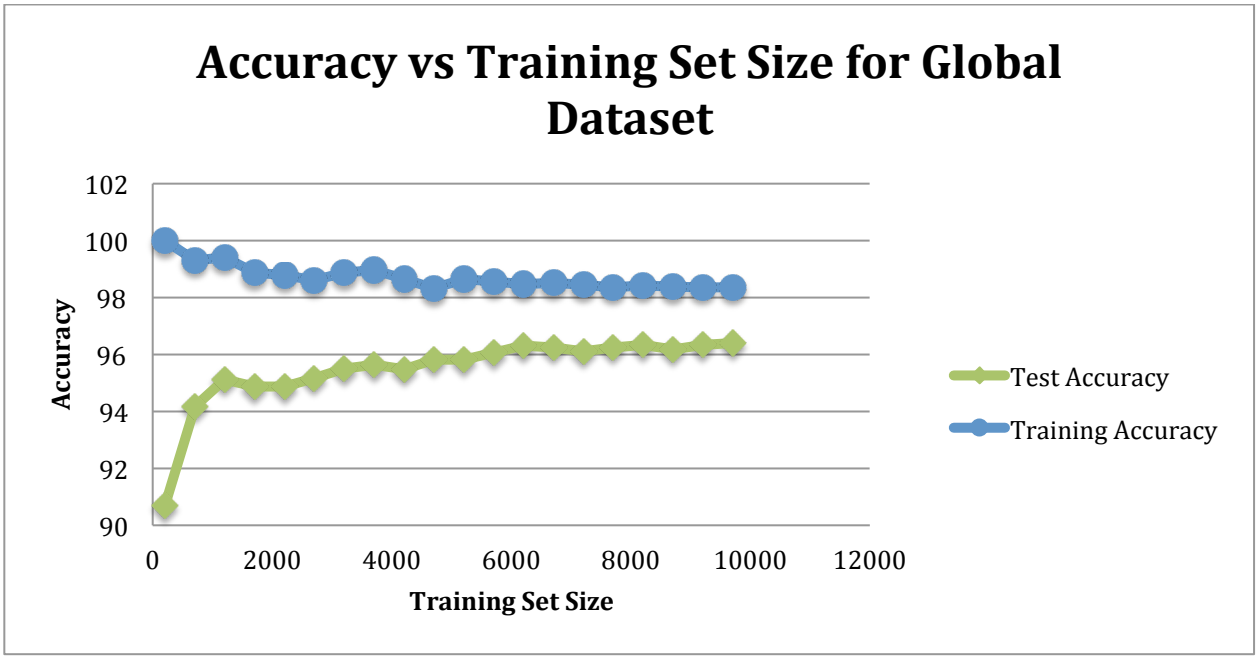


Figure 1 Adding training data reveals minor underfitting (bias) suggesting more features could improve accuracy

For the purpose of eliminating artifacts from the correlation results, we decided that 95% accuracy provided sufficient confidence for deploying the system in a production environment. To demonstrate that our results were meaningful (i.e., that our accuracy was truly accurate), a new capability was added to the manual classification program that, for a given sample, used the trained SVM to predict the quality and display the result to the operator. This prediction tool proved to be a valuable, informal metric for evaluating the classifier.

Classification of the global correlated dataset resulted in the removal of 11,730,663 correlation pairs from the table (i.e., over 70% of the correlations were determined to have at least one bad segment). Eliminating so much invalid data would certainly increase productivity for analysts attempting to sift through the correlation results, but ideally the automated classifier could be used earlier in the processing pipeline to flag non-seismic events before ever attempting correlation. In order to be effective as part of this detection process, the automated classifier would need to generalize well to previously unseen arrays of seismic sensors, or at least require minimal modification beyond gathering a few thousand manual classifications and retraining.

Forward selection was applied to the feature set as a means of informally assessing how much each feature was contributing to the classifier's performance, and to see if any features captured redundant qualities of the data. Using AVG_DELTA alone, the classifier was able to achieve about 78% accuracy, but given that the dataset was skewed 67% towards bad examples, this wasn't particularly remarkable. The selection of AVG_DELTA (a measure of the distance between the event and the sensor) as the most informative feature raised questions about this approach's ability to generalize to processing unseen seismic arrays in real time. Due to being a non-

seismic feature of the station and event, AVG_DELTA would have little meaning for seismic arrays processed individually.

Automatic Classification – Local Dataset

To test the viability of applying the SVM approach as part of a detection pipeline, over ten thousand segments from an entirely new dataset were manually labeled. Unlike the global dataset used in correlation, these data were all local to a single seismic array. Directly applying the SVM trained on the global dataset to the local resulted in dismally poor accuracy of around 60%. Reparameterizing the SVM and retraining on the local labeled dataset increased the accuracy substantially to around 90%. However, inspection of the newly labeled training data revealed that they were skewed 88% towards bad examples. Precision was determined to be upwards of 90%, while recall of good examples was shown to be around 75%. Clearly, the local SVM was achieving high accuracy by blindly throwing out much of the data.

In an attempt to correct this behavior, the local dataset was deskewed for training. More specifically, we selected the largest subset of the manually labeled dataset such that the number of good examples equaled the number of bad examples. This resulted in a training set of approximately 3500 examples. Following reparameterization and retraining of the SVM, the system achieved tenfold cross validation results of 92.41% accuracy, 93.98% precision, and 90.65% recall.

Inspection of how the even subsets of good and bad data were distributed in both the global and local datasets revealed many values differing by orders of magnitude. In an attempt to better capture the similarity between extreme values in different datasets, we logarithmically scaled the feature values. By constraining the data in this way, the parameterization process for both the global and local datasets converged to the same parameters. Consequently, we were able to overcome the limitation of needing to parameterize the SVM for every new dataset (i.e., our selection of the penalty parameter C and kernel parameter γ could remain constant for new seismic arrays, only the training set needed to change).

Conclusions

The same deskewing technique was applied to the global dataset, resulting in an almost 2% increase in tenfold cross validation accuracy to 96.89%. Precision and recall also improved significantly to 96.13% and 97.70%, respectively. The addition of logarithmic scaling and deskewed training data also improved performance when training on the global dataset and testing on the local dataset: accuracy improved by almost 20 percentage points to 79.86%, and precision and recall improved to 73.65% and 92.98%, respectively. Still, the significant hit to performance incurred by not using seismic array-specific training data meant that this approach was not sufficiently accurate for production.

Table 2 Performance metrics with different training and test sets

Test	Accuracy	Precision	Recall	Conclusion
10F CV Global (13866)	96.89	96.13	97.7	Errs on Acceptance
10F CV Local (3446)	92.41	93.98	90.65	Errs on Rejection
Train Global / Test Local	79.86	73.65	92.98	Errs on Acceptance

The difference in tenfold cross-validation accuracy between the deskewed global and local datasets was roughly 4%. Reapplying feature selection to both datasets, we found that our intuition regarding AVG_DELTA was correct. In the deskewed global dataset, AVG_DELTA became a less informative feature, but still contributed significantly to total accuracy (see Figure 2). In the local dataset, however, adding AVG_DELTA to the SVM

actually resulted in a decrease in overall accuracy (-0.09%). We attribute this decline to minor changes in how the SVM fit the training set with AVG_DELTA, resulting in slightly worse generalization to the test set. In both datasets, variance (in the time and frequency domains) proved to be the most valuable discriminator. The window length, signal-to-noise ratio, and positive kurtosis were also consistently valuable features between datasets.

That the majority of features provided little to no gains in feature selection strongly suggests that they captured redundant qualities of the data. In looking to improve performance on seismic array-specific classifiers, we will need to explore additional features.

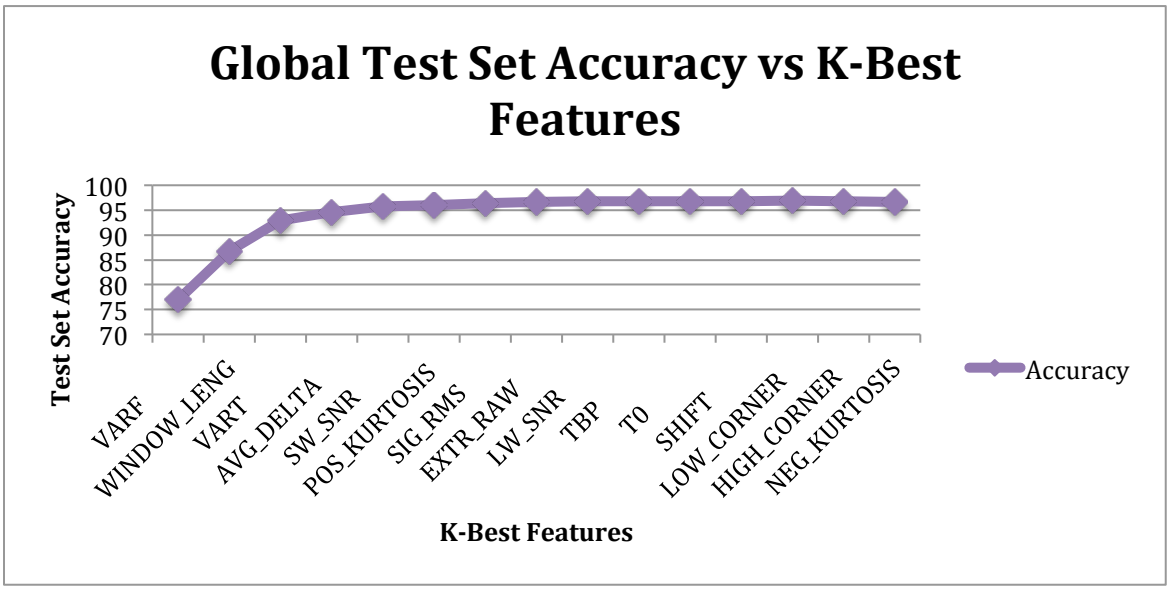


Figure 2 Most accuracy from few features implies several captured redundant qualities of the data

Future Plans

In adopting this automated classification system to an active seismic detection framework, each seismic array will require its own training set and SVM. As new seismic arrays are introduced, a human operator will need to label at least one thousand good and bad examples. After this initial labeling is completed, it is hoped that the automated classifier will be able to run without retraining for months (if not years) at a time. Because the SVM parameterization has been shown to generalize to new datasets, however, the training data will be the only component of the system that will likely need to change between seismic arrays.

There are plans to continue improving upon the capabilities and robustness of the automated classifier. Expanding its use beyond the correlation and detection problems, the classifier could be modified to classify non-seismic data into various artifact subtypes. Additionally, new features could be derived from similar, non-machine learning quality control tools such as IRIS's Quality Analysis Control Kit (QUACK)[3]. More ambitiously, it is hoped that the work here could be developed further to create a new industry standard in seismic quality control and event classification.

Acknowledgements

Douglas Dodge with Lawrence Livermore National Laboratory for providing domain expertise, labeling the training set, generating the unlabeled correlation table, and presenting the original problem.

[3] http://www.iris.edu/dms/newsletter/vol8/no1/exploring-iris-data-and-data-quality-with-quack/