Application of Supervised Machine Learning Classification Techniques for Preprocessing Passive Seismic Earthquake Data

Nicholas Appiah[1*] and Harold Gurrola[1]

[1] Department of Geosciences, Texas Tech University, Lubbock TX

* Corresponding email: nappiah@ttu.edu

Abstract

P-Receiver Functions (P-RFs) often employ the analysis of hundreds of thousands of seismograms to passively image Earth's crust and upper mantle structure. This study compared the performance of 6 supervised machine learning classification techniques in quality checking seismograms. Quality checking seismograms prior to computing P-RFs is very vital in distinguishing between usable ones and non-usable ones. 372,087 seismograms covering the region of California were high-pass filtered and low-pass filtered at 0.02 Hz and 0.5 Hz respectively after which 8 features were computed. All the seismograms were then quality-checked manually to generate labeled data. The labeled data was partitioned into training set and test set which were used to train and evaluate the 6 classification techniques, respectively. Out of the 6 techniques evaluated, the Decision Tree had the highest accuracy of 97.2% and therefore had the best performance. The accuracy of all other techniques was above 94% except the Discriminant Analysis which had a low accuracy of 75.2%. The Decision Tree, however, had much more trouble classifying non-usable seismograms correctly than the usable ones due to overlap in features between the two response classes.

Introduction

- P-Receiver Functions (P-RFs) are routinely used to passively image crust and upper mantle structure by deconvolving the horizontal component by the vertical component of hundreds of thousands of seismograms to isolate the P-to-S converted phases.
- Prior to computing the P-RFs, as a preprocessing step, it is crucial to quality check the seismograms to ascertain that the desired earthquake signal (the first P-arrival in this case) is not masked by background noise. By so doing, we can distinguish between usable events and non-usable events (see figure 1).
- Quality checking the hundreds of thousands of seismograms has traditionally been done manually by visually inspecting each of the seismograms, one at a time, to determine the usable events. However, this approach is very laborious, time-consuming, and may be prone to human errors.
- In this study, the goal is to employ machine learning (ML) to automate the process of quality checking the seismograms. Specifically, we seek to classify the seismograms into usable and non-usable using 6 supervised ML classification techniques. We also seek to compare the performance of the techniques to determine which works best.

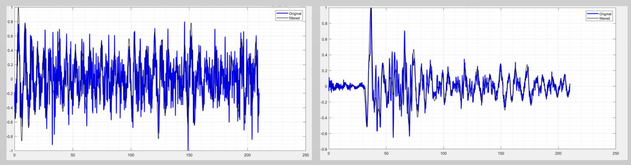

Figure 1: A vertical component seismogram in which the desired earthquake signal (first P-arrival) is (left) masked by background noise and (right) not masked by background noise. The seismogram on the right is usable for P-RF analysis while the one the left is not. Signal time window is between 29 and 40 seconds whereas noise time window is between 5 and 20 seconds. In blue is the original unfiltered seismogram whereas in black, the seismogram is filtered between 0,02 Hz and 0.5 Hz.

Methodology

Data

The data consists of 372,087 seismograms from earthquakes of magnitude 5.8 and above and great circle arc of 30° – 90° between event and station. The seismograms are from 1990 to 2021 and cover the region of California between latitudes 31° and 42° N and longitudes 113° and 125° W.

Feature Extraction and Generation

Both peak signal-to-noise ratio and standard deviation of signal-to-noise ratio were computed from seismograms high pass filtered at 0.02 HZ (StoN_peak_high, StoN_std_high) and then low pass filtered at 0.5 Hz (StoN_peak_low, StoN_std_low). Four (4) additional features, tot_SN, tot_SN_high, tot_SN_low, and tot_SN_peak were computed. Figure 2 shows the distribution of the features.

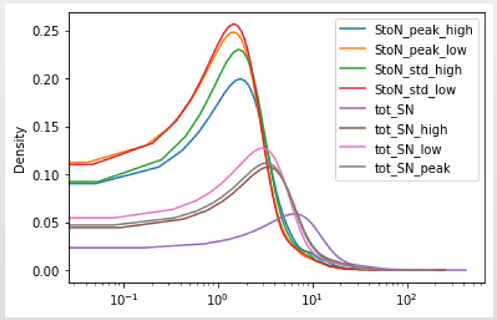

Figure 2: A Kernel Density Estimation plot showing the distribution of all 8 features. Values for all features are mostly centered between 0.5 and 10 with the majority being centered at 1.

Generation of Labeled Data

All the 372,087 seismograms were quality-checked manually to generate labeled data. The usable and non-usable seismograms were designated a response class output of 1 and 0, respectively. Figure 3 shows a coordinate plot of the features.

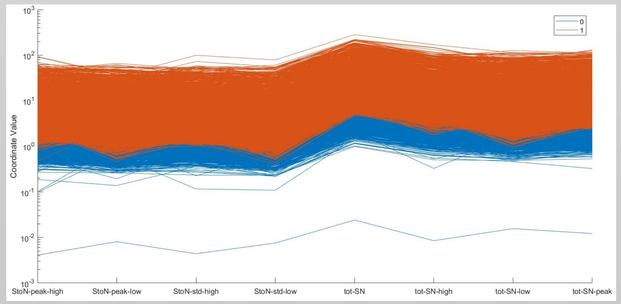

Figure 3: A coordinate plot of the features grouped by the response class output. From the plot, there is always some amount of overlap between features of the two response output classes. That means, although the two classes can be separated, there is bound to be some confusion i.e., the machine learning techniques are bound to be confused when doing classification in the area/region of overlap .

Partitioning of Labeled Data into Training Set and Test Set

The labeled data was partitioned into training set and test in the ratio 7:3, respectively. The training set and test set were used to train and evaluate the ML techniques, respectively. The features of both training set and test set were scaled to a mean of 0 and a standard deviation of 1 prior to training and evaluating the ML techniques.

Classification Techniques

Six (6) supervised ML classification techniques (figure 4) were trained using the training set. These are: K Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB), Discriminant Analysis (DA), Support Vector Machine (SVM), and Artificial Neural Network (ANN)

The KNN uses a group of K seismograms, known as the nearest neighbors, whose classes are known, to classify a seismogram whose class is unknown. The optimum value of K was found to be 1. The DT classifies an unknown seismogram by choosing the best possible split iteratively for each feature based on a given criterion until no further splits can improve the criterion. Both the NB and DA assume that the seismograms in each response class are statistical samples from normal probability distributions. While the NB classifies an unknown seismogram by computing the probability that it comes from a given response class, the DA does so by determining the location of a boundary between the response classes where probabilities are equal. The SVM classifies an unknown seismogram by finding the best hyperplane that separates the response classes in the feature space. The ANN classifies an unknown seismogram by iteratively adjusting the connections (the weight value) between neurons in the hidden layer(s) through trial and error. The ANN used in this study consisted of 3 hidden layers each containing 10, 12 and 10 neurons.

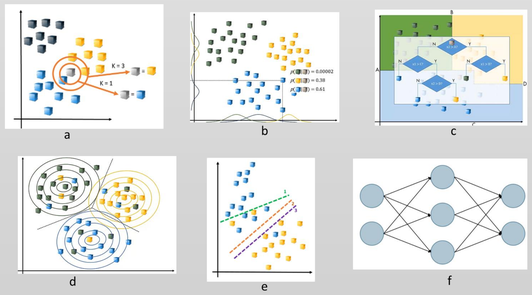

Figure 4: Supervised ML classification techniques employed in this study; (a) KNN (b) NB (c) DT (d) DA (e) SVM and (f) ANN

Results and Discussions

Evaluation of Classification Techniques

The 6 ML classification techniques were evaluated using the test set (figures 5, 6, and 7). From figure 5, DT had the highest accuracy of 97.2% and had the best performance. This was followed by ANN (97.0%), SVM (96.6%), KNN (96.5%), NB (94.8%) and DA (75.2%) which had a low accuracy. From figure 6, The false negative rate for response classes 0 and 1 are 5% and 3.6% respectively. This implies that the DT had more trouble classifying non-usable seismograms than the usable ones. From figure 7, It is evident that the misclassification of non-usable seismograms occurred due to overlap in features between the two response classes; the features are not able to distinguish between the two classes. Hence, engineering new features or using a different classifier in an ensemble manner may be a probable solution to resolve the confusion.

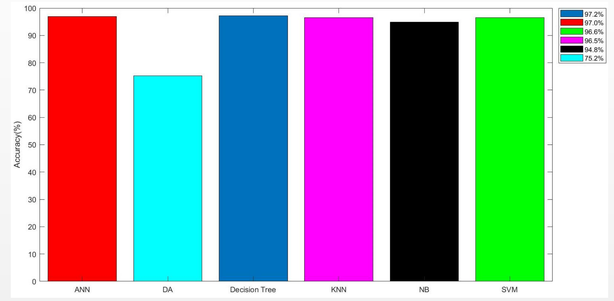

Figure 5: Bar chart showing the accuracy of the 6 classification techniques.

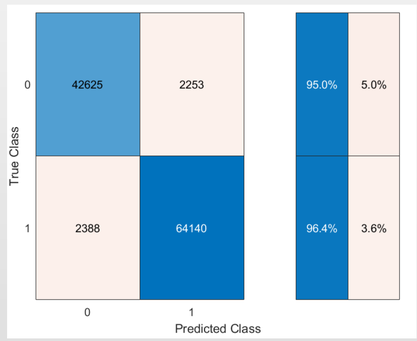

Figure 6: Confusion matrix for the DT, the best technique.

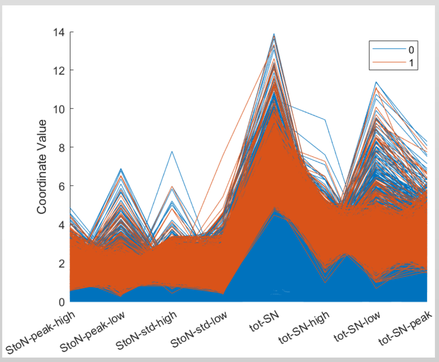

Figure 7: Coordinate plot for true positive non-usable seismograms and false negative non-usable seismograms. These represent non-usable seismograms (0) that were correctly classified as non-usable seismograms (0) and non-usable seismograms (0) that were misclassified as usable seismograms (1), respectively.

Conclusion and Recommendations

The performance of 6 different supervised machine learning classification techniques for quality checking seismograms were evaluated and compared. The Decision Tree classifier gave the best results with overall accuracy of 97.2%. To improve the accuracy of the Decision Tree, it is recommended that (i) new features are computed and (ii) different classifiers are used in an ensemble manner for the seismograms where misclassification occurred.