Data-Driven Earthquake Location Method
Project Report

Weiqiang Zhu (06118474), Kaiwen Wang (06122739)
Department of Geophysics, School of Earth, Energy and Environmental Science

12/16/2016

1 Abstract

Earthquake location is one of the most fundamental and important problems in Geophysics. Traditional earthquake location methods depend on dense array and high signal to noise rate of data. In this project, we try to apply machine learning algorithms trained on historical seismic wave records to locate earthquakes which only needs one or few stations and independent of prior physical knowledge.

2 Introduction

Earthquake is one of the most commonly happened disasters with devastating consequences and poses significant risks to human society. The ability to determine the time and location of an earthquake can provide valuable information for rescue guidance, hazard assessment and scientific study. After a destructive earthquake, an accurate earthquake location is the most importtance information needed for rescue. Earthquakes also provide valuable information for scientists to study our Earth: the movement of plates, the property of rock in Earth's interior and local geological background.

The traditional way to get earthquake locations is to manually pick the first arrival times of seismic waves and run inversion algorithms based on seismic-wave velocity model of Earth (Figure 1).

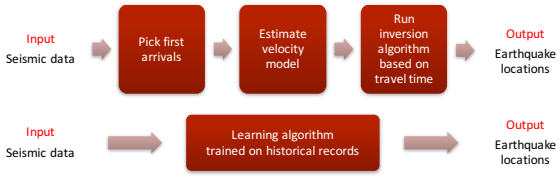


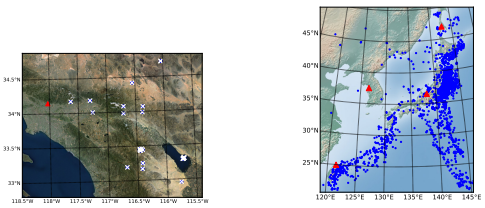
Figure 1: Earthuqake location methods

In this project, we tested two machine leaning algorithms to do earthquake locating: k-nearest neighbors(k-NN) algorithm and convolutional neural network (CNN) algorithm. The inputs of our data set are seismograms (the waveforms of ground shaking) of earthquakes recorded at seismic stations. The outputs are the latitudes and longitudes of the epicenters of these earthquakes. We expect the machine learning methods to capture more features using the full waveforms of earthquakes, instead of only using the first arrival times and figure out the relationship between these features to predict earthquake locations.

3 Related Work

One of the current research direction of improvement of earthquake locations is to deploy a dense array with bore hole stations (Hi-net in Japan) [1, 2]. The dense array provides more

data and the bore hole stations have higher signal to noice ratio. Since the cost of deploying and maintaining high resolution seismometers are relatively high, it's not practically possible to cover the Earth with dense bore hole array. Another approach is the Double Difference method [3], which locate earthquakes using relative locations. This method could reduce the error based on a reference event. However, the events must be close enough(within $\frac{1}{2}\lambda$), and the error of relocated events depends on the error of the reference event. Our goal is to apply machine learning algorithms to locate earthquakes with one or few stations using more and more hitorical earthquake location records.



(a) Dataset in California (b) Dataset in Japan

Figure 2: Seismic stations

4 Dataset and Features

Two datasets are used here: one in California and the other in Japan. For the CA dataset, we use earthquakes between 5/11/2016 and 11/11/2016 in Southern California (latitude range: 34.3~39.3N, longitude range: 135.7~145.7E). The waveforms are recorded at a seismic station whose Network code is CI and station code is Q0001. In total, there are 30 events (Figure 2(a)). We use each of them as test set while the other 29 as training set. For the Japan dataset, we use earthquakes whose magnitude above Mw 5 between 01/01/1998 and 10/06/2016 in east asia (latitude range: 21.0~51.0N, longitude range: 117.1~147.1E). Four seismic stations in Network IU of codes: MAJO, YSS, INCN and TATO are used. In total, there are 2510 events (Figure 2(b)). We use the first 100 data as validation set and the rest as training set.

Figure 3 shows examples of the seismic records. Figure 3(a) is a record with high signal to noise ratio. Figure 3(b) is a record with low signal to noise ratio. We can see clearly the waveform generated by an earthquake. Figure 3(c) and 3(d) shows Hilbert transformation of the good and bad data. Figure 3(e) and 3(f) shows Wavelet transformation of the good and bad data. All records regardless of high or low signal to noise ratio are used in training at current stage.

5 Methods

5.1 k-NN

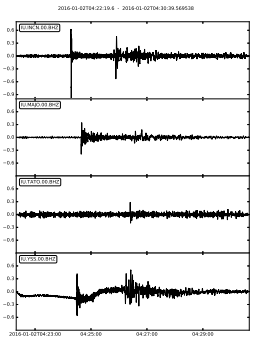
k-nearest neighbors algorithm(k-NN) is first tested to predict earthquake locations. The method is based on the fact that nearby earthquakes are similar in waveforms. The similarity between two waveforms are expressed in correlation distance. The prediction of location is an average of the location of k nearest neighbors, weighted by the inverse of their correlation distance to the query event.

This algorithm works as follows:

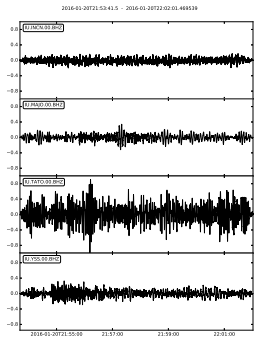
Step 1. Compute the distance in waveforms from the test example to training examples.

Step 2. Find the k-nearest neighbors of the test example.

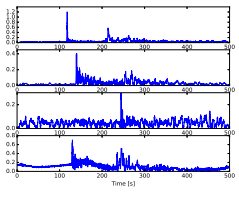
Step 3. Calculate an inverse distance weighted average of the k-nearest neighbors.



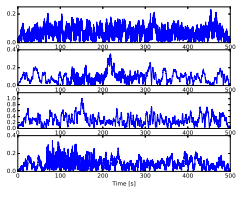
(a) Good raw data



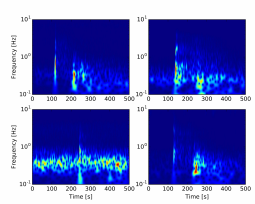
(b) Bad raw data



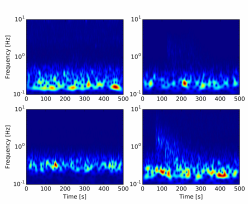
(c) Hilbert transformation of good data



(d) Hilbert transformation of bad data



(e) Wavelet transformation of good data



(f) Wavelet transformation of bad data

Figure 3: Examples of training set data

5.2 Convolutional neural network

Convolutional neural network (CNN) [4] is shift invariant and good at understaning the spatial relationship inside a picutre. The seismic waveform (Figure 3(a)) consists of many different features like: body waves, surface waves, P wave, S wave and many other phases [5]. All these phases have different speeds and arrive at specific times at different stations, so they have rich information about the source locations. If CNN can learn these features from seimic waveforms, it may be able to predcit the earthquake locations.

6 Results

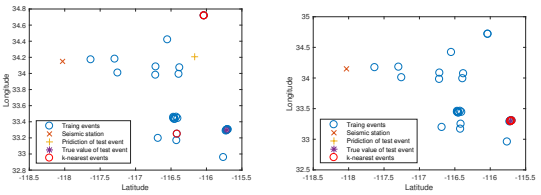
6.1 k-NN

First we try to run k-NN regression quick-and-dirty on raw data to get a main idea of what's happening. We randomly chose an event to be the query event from out set of 30 events. The regression result is in Figure 4(a). The mean error is shown in Table 1.We could see clearly that finding better nearest neighbors account for most of the improvement. That is, our next step should focus on improving the accuracy when calculating the distance between waveforms. There are mainly two ways to achieve that. One way is to process the input waveforms and the other way is to choose a better distance function.

We first tried several ways to process our input waveforms. We tried removing the trend and mean, tapering, filtering with different frequency band and Hilbert transformation. The best condition to process data is to remove the mean and trend, taper, filter between 0.1 to 5 Hz, Hilbert transform and low pass filter under 2 Hz. The regression result of the processed data is shown in Figure 4(b).

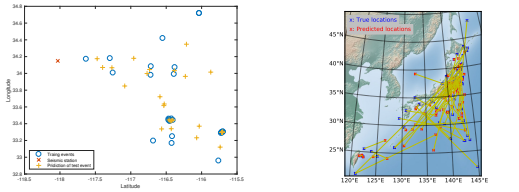
Then we tried to choose a better distance function. We found that correlation distance has smaller error compared with Euclidean distance. We also chose a better k and further reduced the mean error. The result is shown in Figure 5(a). Using three-components data did not give a smaller error. Possibly because the Z component has better records of local earthquakes compared with the other two components.

We also applied the method to Japan dataset. The result is shown in Figure5(a). The mean error is shown in Table 1. For the same reason as using multiple components, using multiple stations also did not reduce the error.



(a) Input raw data (b) Input processed data

Figure 4: Plots for prediction of a query event



(a) CA dataset (b) Japan dataset

Figure 5: Predictions of CA and Japan dataset

6.2 CNN

6.2.1 CNN structure

The stucture of our CNN is show in Figure 6. The neural network consisits of two covolutional layers, two max pooling layers and a fully connected layer with 50 percent drop out. The nurel network is trained using 2510 earthquakes around Japan. The first 100 samples are used as the validation data set. The mini-batch size is setted as 100. The start learning rate is 0.01 and the decay rate is 0.99. Our CNN is implemented based on Tensorflow [6].

Table 1: mean error of different improvements

Improvements	Evlo	Evla
Train on raw data(CA)	0.4066	0.5250
Data processing(CA)	0.2201	0.3822
Distance function(CA)	0.1351	0.2654
Parameter k(CA)	0.1371	0.2396
Three-components(CA)	0.4177	0.3313
CA data location limit	0.0342	0.0530
More events(Japan)	2.8803	2.9646
Four stations(Japan)	2.9505	3.8858
Japan data location limit	0.0601	0.0623

The input data are the raw data, data after hilbert transform and data after wavelet transform. Because 4 stations are used in this paper, the input data consist of four channels. Due to limitiation of computatioal power, we have to apply down-sampling of the origial data. The data shape of each channle is 1×1000 for raw data and data after hilbert transform, 25×50 for data after wavelet tranform.

The output data are predicted coordinates: logitudes and latitudes. L2-norm of difference between true and predicted coordinates is used as the loss function.

6.2.2 Predicted results of CNN

The training errors using data after wavelet transform is shown in Figure 7. The predicted

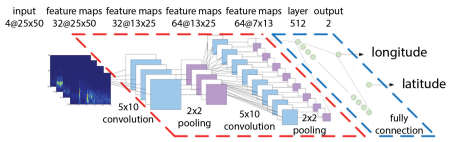


Figure 6: Convolutional neural network structure

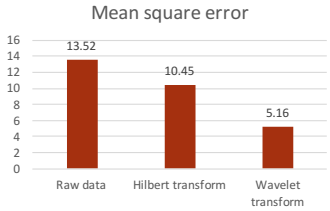


Figure 9: Training errors of different input data types

earthquake locations of the validation data set is shown in Figure 8. Most of the predicted locations are very near to their true locations. The CNN method achieves a very high prediction accuracy.

The comparison of training erros between raw data, data after hilbert transform and data after wavelet transform are shown in Figure 9. The result of data after wavelet transform shows the lowest training error on valid data set.

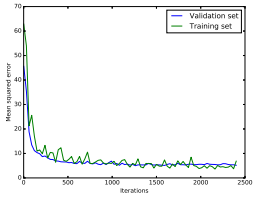


Figure 7: Training errors of data after wavelet transform

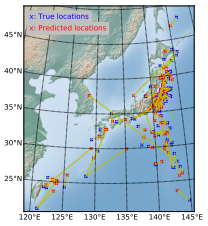


Figure 8: Predicted earthquake locations

7 Conclusion and Future Work

In this report, we have successfully applied mathine learning methods to earthquake locations,which has never been tried before. Both algorithms (kNN and CNN) show reasonable prediction results. CNN has better accuracy while testing on the data in Japan. But currently the two algorithms could not provide as accurate results as traditional methods. Possible reasons for mislocating may include: 1. background noise 2. not enough historical data. 3. errors of finding nearest neighbor. 4. errors in down sampling of input data. Further improvements would be required to put them into practical application.

To find out the main reason for high error of kNN, we applied error analyses. Due to the limit of events catalog, the perfect result(if we find all the k-neighbor correctly) should have a mean error of (0.03,0.05)(shown in Table 1), which is still a magnitude lower to current error. Thus, for future research, the most room for improvement is still finding k-nearest neighbor more precisely.

For CNN, higher data sampling rate and deeper neural network may help to improve the accuracy furture as it can capture more accurate information of different arrival times of phases.

References

[1] Yoshimitsu Okada, Keiji Kasahara, Sadaki Hori, Kazushige Obara, Shoji Sekiguchi, Hiroyuki Fujiwara, and Akira Yamamoto. Recent progress of seismic observation networks in japanhi-net, f-net, k-net and kik-net. *Earth, Planets and Space*, 56(8):xv–xxviii, 2004.

[2] Kazushige Obara, Keiji Kasahara, Sadaki Hori, and Yoshimitsu Okada. A densely distributed high-sensitivity seismograph network in japan: Hi-net by national research institute for earth science and disasterprevention. *Review of scientific instruments*, 76(2):021301, 2005.

[3] Felix Waldhauser and William L Ellsworth. A double-difference earthquake location algorithm: Method and application to the northern hayward fault, california. *Bulletin of the Seismological Society of America*, 90(6):1353–1368, 2000.

[4] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[5] Peter M Shearer. *Introduction to seismology*. Cambridge University Press, 2009.

[6] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.