Developing Ground Motion Prediction Equation (GMPE) by Machine Learning

PAN Mengxin
Department of Civil Engineering
HKUST

Abstract

Ground Motion Prediction Equation (GMPE) plays an important role in earthquake hazard mitigation. It is a set of empirical equations evaluating the ground motion for a typical site based on the earthquake magnitude, distance, depth and other features during the earthquake. The drawback of traditional GMPE is that the functional forms had been pre-defined, which means that it is hard to find the optimal model. In this study, the Neural Network, SVM, and random forest have been employed to do the experiments. The results got by machine learning models are better but not much better than the results got by the traditional models.

1 introduction

Earthquake is inevitable and it can not be predicted so far. However, a lot of researches are conducted to mitigate the earthquake risk around the world. When estimating the damage of the earthquake to a city, people rely on the ground motion for a typical site instead of earthquake magnitude, so Ground Motion Prediction Equations (GMPE) plays a critical role in the earthquake risk mitigation. It evaluates the ground motion of a site, such as peak ground acceleration (PGA), based on the earthquake magnitude, distance from source to site, depth, site conditions and other parameters.

Due to the high nonlinearity and large uncertainty of the earthquake data, current GMPE developed by traditional regression methods are not accurate enough. Machine learning has extraordinary ability to extract the characteristic from different datasets. It contains many advanced algorithms to do prediction and regression. Theoretically, it is possible to develop GMPE with smaller errors by machine learning.

The slope of this study is to compare the results of machine learning models with a famous traditional GMPE named (CB14) developed by two great researchers Campbell and Bozorgnia in 2013.

2 Database

The ground motion database used in this study is a subset of the PEER NGA-West2 Database. It was completed in 2014 and it is the largest and the most completed earthquake database in the world. The NGA-West2 database includes over 21,000 three-component recordings from California and worldwide earthquakes with moment magnitudes (M) ranging from 3.0 to 7.9. After applying general exclusion criteria, 15493 recordings from 322 earthquakes were employed in this study.

There are 10 features/inputs and 1 output in this regression model. The output is the Peak Ground Acceleration (PGA) of the typical location. The definitions of inputs are defined as follows:

M is moment magnitude. R_RUP (km) is closest distance to the coseismic fault rupture plane (a.k.a. rupture distance). R_JB (km) is closest distance to the surface projection of the coseismic fault rupture plane (a.k.a. Joyner-Boore distance). R_X (km) is closest distance to the surface projection of

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

the top edge of the coseismic fault rupture plane measured perpendicular to its average strike. W(km) is the down-dip width of the fault rupture plane. λ ($^\circ$) is rake angle defined as the average angle of slip measured in the plane of rupture between the strike direction and the slip vector. Z_TOR (km) is the depth to the top of the fault rupture plane. δ ($^\circ$)is the average dip angle of the fault rupture plane measured from horizontal. V_S30 (m/s) is the time-averaged shear wave velocity in the top 30 m of the site (a.k.a. 30 m shear wave velocity). Z_HPY(km) is the hypocentral depth of the earthquake measured from sea level.

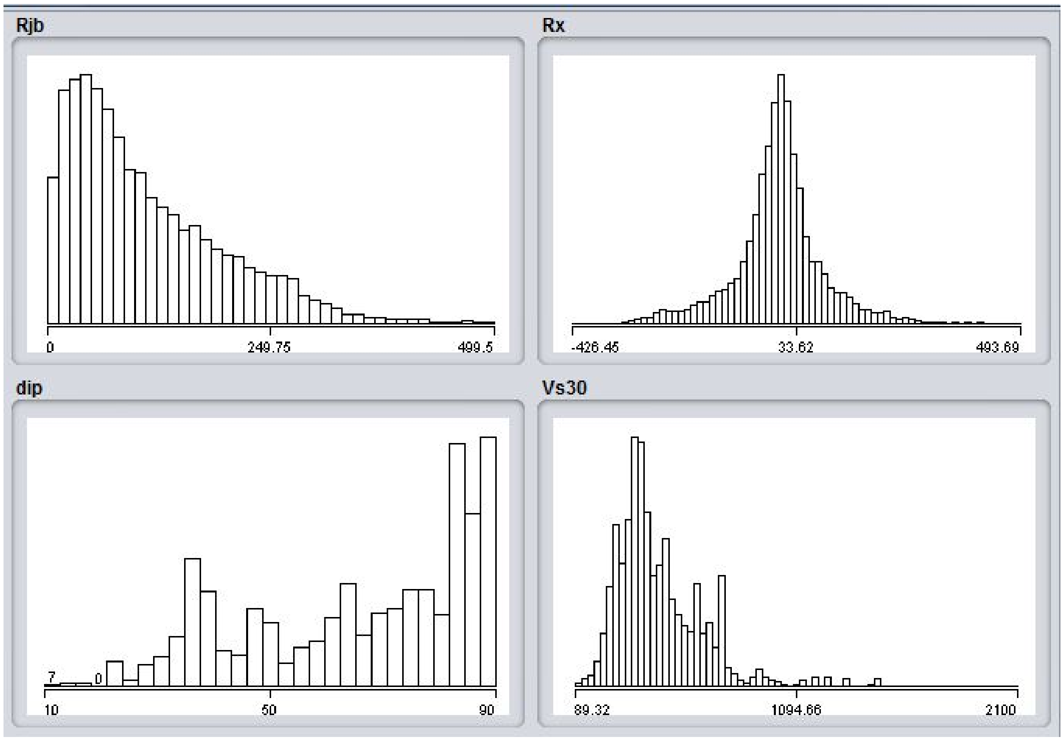


Figure 1: visualization of input data

In addition, most of the PGAs are between 0 and 0.1. In order to disperse the PGA values, I got the logarithm of PGA and get a better distribution. Both PGA and lnPGA will be used to do the regression.

3 Current Ground Motion Prediction Equation (GPME)

The ground motion prediction equation named CB14 developed by Campbell and Bozorgnia in 2014, which is one of the most famous GMPE in the world. The functional forms of CB14 are pre-defined based on the previous observation and simulation. All the data were used to train the model. since the model is simple, containing only about 30 parameters, overfitting would not happen. People do not need to divide the validation or testing set from the total data. For all the 15493 records,when the output is PGA, the mean square error is 0.001988 and the correlation coefficient is 0.8409. When the output is lnPGA, the MSE is 0.6506 and the R value is 0.9470.

The current GMPE seems good, but the accuracy still needs to be improved. The features used to train the machine learning models are consistent with the current model (CB14). Using the same database and the same features in different models seems like a robust way to do the comparison.

4 Neural Network

Both feedforward network and cascade-forward network were employed in this study. Different training functions were used in this study, such as Levenberg-Marquardt (LM), BFGS Quasi-Newton,

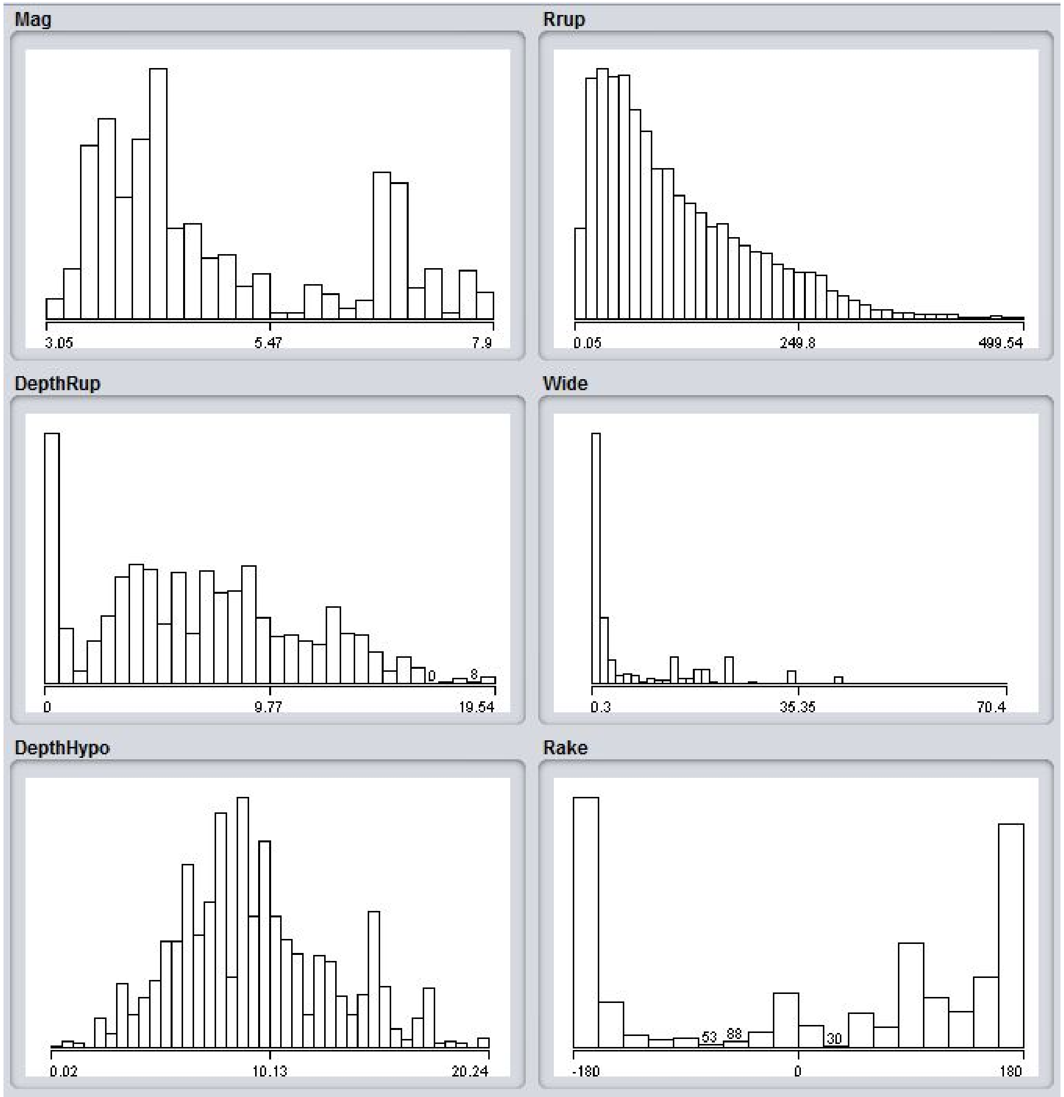


Figure 2: visualization of input data

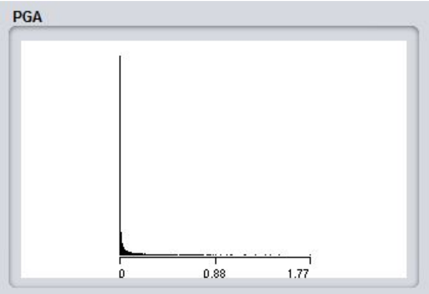


Figure 3: visualization of output(PGA)

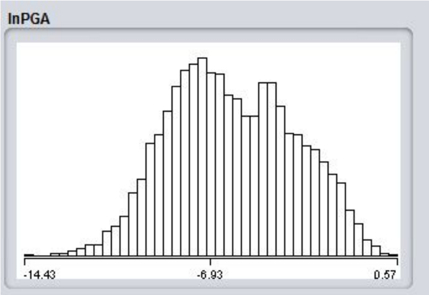


Figure 4: visualization of output(lnPGA)

Resilient Backpropagation (RP) and Scaled Conjugate Gradient (SCG).Levenberg-Marquardt is the best method in this study instead of other more complicated methods. In addition, the training functions have more effect in the convergent speed instead of the model accuracy. Since the dataset is not very large in this study, decreasing the convergent speed is not the scope herein.

The networks with different numbers of hidden layers and neurons in each hidden layers were also conducted in this study. The results of neural network models are not stable enough, since the simple neural network can only find the local optimal instead of the global optimal, which means that neural network is sensitive to the initial weights and bias. Every model was trained fine times with different random initial weights and bias in this study, and then the one with the smallest mean square error in the validation set was chosen. However, from the results shown in figure 5 to figure 11, the results still not stable enough, especially for the testing set.

In 1 layer model, the increase of number of neurons would not lead to the decrease of MSE. As for 2-layer and 3-layer models, the MSE of training set and validation set decreased a little, but the MSE of testing set became larger, which means that overfitting appeared in this situation. The network with 1 hidden layer and 6 neurons has already get the similar accuracy which more complicated feedforward network can get. The optimal feedforward model with the smallest MSE in validation set contains 1 hidden layer and 24 neurons. The MSE in training set, validation set and testing set are 0.001508, 0.001271, 0.001517, respectively.

Cascade-forward networks are similar to feed-forward networks, but include a connection from the input and every previous layer to following layers. The cascade-forward networks with 3 hidden layers had a good performance. The optimal feedforward model with the smallest MSE in validation set contains 3 hidden layers and 30 neurons in each layers. The MSE in training set, validation set and testing set are 0.001578, 0.001065, 0.001491, respectively.

In conclusion, although neural network can get smaller MSE than the current GMPE (CB14), the results are not stable enough, so the feedforward or cascade-forward neural networks are not good enough in this regression study.

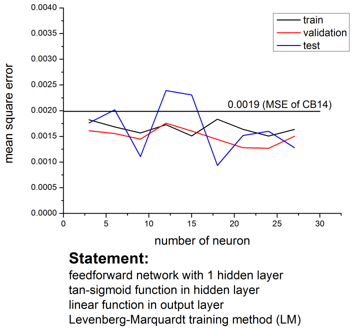


Figure 5: feedforward network with 1 hidden layer(LM)

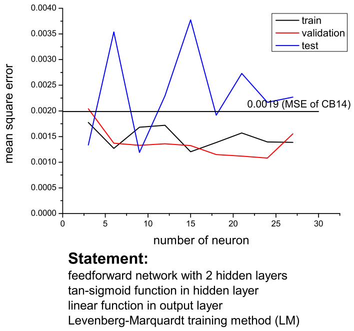


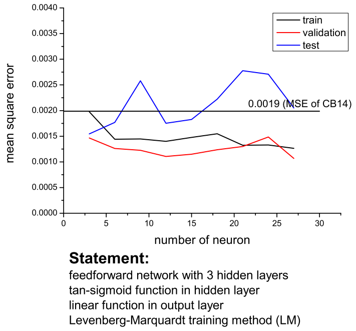
Figure 6: feedforward network with 2 hidden layers(LM)

5 Support Vector Machine (SVM)

In this section, the LIBSVM was utilized to train the model. LIBSVM is an advanced SVM library developed by Prof. Chih-Jen Lin from National Taiwan University.

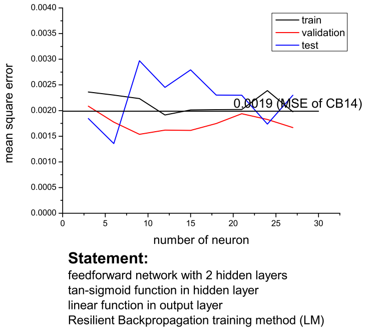
Firstly, the PGA without logarithm was used to train the LIBSVM model. The inputs were normalized between 0 to 1. All the kernel functions were employed to conduct the model, like linear, polynomial, Gaussian, and sigmoid. I also conduct the grid-search method to adjust the c and g parameters to find the optimal model. However, it seems that none of them could get the good results. The linear kernel had low accuracy, while the Gaussian kernel leaded to overfitting easily in this situation. The smallest MSE got by SVM with Gaussian kernel function is 0.0019, and R is 0.8582.

when regarding the logarithm of PGA as the output, the results became better. Also, I used grid-search method to find the best c and g parameter when employing Gaussian kernel. The best model was



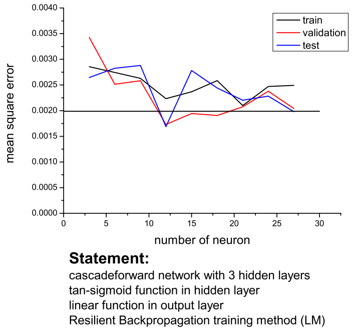
Statement:
feedforward network with 3 hidden layers
tan-sigmoid function in hidden layer
linear function in output layer
Levenberg-Marquardt training method (LM)

Figure 7: feedforward network with 3 hidden layers(LM)



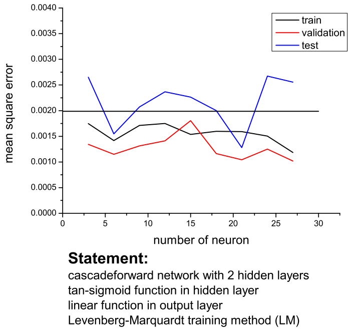
Statement:
feedforward network with 2 hidden layers
tan-sigmoid function in hidden layer
linear function in output layer
Resilient Backpropagation training method (LM)

Figure 8: feedforward network with 2 hidden layers(RP)



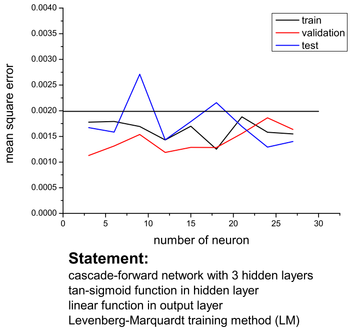
Statement:
cascadeforward network with 3 hidden layers
tan-sigmoid function in hidden layer
linear function in output layer
Resilient Backpropagation training method (LM)

Figure 9: feedforward network with 3 hidden layers(RP)



Statement:
cascadeforward network with 2 hidden layers
tan-sigmoid function in hidden layer
linear function in output layer
Levenberg-Marquardt training method (LM)

Figure 10: cascade-forward network with 2 hidden layers



Statement:
cascade-forward network with 3 hidden layers
tan-sigmoid function in hidden layer
linear function in output layer
Levenberg-Marquardt training method (LM)

Figure 11: cascade-forward network with 2 hidden layers

output	PGA		lnPGA		lnPGA back to PGA	
criteria	MSE	R	MSE	R	MSE	R
CB14	0.001988	0.8409	0.6506	0.947	0.001988	0.8409
Feedforward NN	0.001271	0.8673	0.4619	0.953	0.002519	0.7728
Cascade-forward NN	0.001065	0.8698	0.4392	0.9635	0.010846	0.5427
SVM (Gaussian kernel)	0.0019	0.8582	0.5139	0.9587	0.002	0.8554
Random forest	0.001689	0.8646	0.4354	0.9647	0.00239	0.6845

Table 1: The comparison of different models (validation set)

found when c and g are both equal to 3, and the MSE in validation set is 0.5139 and the R value is 0.9587.

6 Random Forest

I also employed the random forest method in Weka. When the output is PGA, the best MSE is equal to 0.001689 and R is 0.8648. When the output is lnPGA, the best MSE is 0.4354 and the R is 0.9647. It seems that random forest is also an optional method in this study.

7 Conclusion

The comparison of R and MSE for the validation set is shown in table 1.In the last two columns, based on the predicted lnPGA, the predicted PGA can be calculated. Based on the new predicted PGA, the MSE and R can be calculated. When the output is PGA, cascade-forward network can get the best result. When the output is lnPGA, the random forest can get the best result. The result from different machine learning models are not much better than the traditional model (CB14). Although when the output is lnPGA, the R can be about 0.95, but the goal of the Ground Motion Prediction Equation (GMPE) is to get the predicted PGA instead of lnPGA.As shown in the last two columns in table 1, the MSEs are quite large and the R is pretty small. A little bias in lnPGA will lead to a big bias in PGA, so when the models built by lnPGA were still not accurate enough.

8 Discussion

The main drawback of the models conducted in this study are low accuracy and instability. There are some solutions I want to try in the future. Firstly, the data pre-processing is very important. The distributions of output in this study is so wired. Some of the input features are similar to each other, such as different kinds of source-to-site distance, maybe some clustering can be conducted herein. Secondly, more robust method needs to be used in this situation to solve the instability problem.