15th World Congress on Computational Mechanics (WCCM-XV)
8th Asian Pacific Congress on Computational Mechanics (APCOM-VIII)
Virtual Congress: 31 July – 5 August 2022
S. Koshizuka (Ed.)

# Data augmentation technique for construction engineering regression surrogate model

**KAI OGATA[1], YOSHITAKA WADA[2]**

[1]Kindai University
Higashi-Osaka Japan
2133330319m@kindai.ac.jp

[2] Kindai University
Higashi-Osaka Japan
2133330319m@kindai.ac.jp

**KEY WORDS:** Surrogate model, Data Augmentation, Convolutional Neural Network, Deep Learning

**Abstract.** *The objective of this study is to predict the degree of danger to the human body from motion information such as acceleration, velocity and displacement during a collision between a car and a human body. As a preliminary step, the maximum bending moment that occurs in the leg was predicted using a convolutional neural network. The responses which are represented by learning data generated by 1D-CAE system. A number of training data sets are varied in order to show the enough number to predict. The predictor's accuracy is evaluated by the test data sets. We'd like to discuss necessisty of a total number of training data sets and effectiveness of data augmentation technique. In addition, the technique to utilize classification by the t-SNE method to improve accuracy is also examined. t-SNE is based on classification algorithm, however an engineering interpolation should be computed based on physical meanings and influential parameters.*

## 1 INTRODUCTION

In these days, simulation-based verification methods are utilized more frequently rather than verification through experiments using the product in the field of industrial product design. In this context, a development of an evaluation method instead of CAE is needed for more effective designing of industrial products. Machine learning methodology is one of the most anticipated way to accelerate evaluation process.An application of a surrogate model in deep learning to engineering problems is still studied. The accuracy of the prediction is improved by augmentation technique, which is called as oversampling to avoid overfitting. The augmentation is conducted using data sets with low prediction accuracy.

## 2 EXPERIMENTAL DETAILS AND DATA AUGMENTATION

### 2.1 Least Squares and Maximum Likelihood Methods

In this study, a regression predicts the output y from the input x. We assume that the output is a scalar and is a parametric model. y is always represented by the following equation.

$$\hat{y} = f(x; w) \tag{1}$$

However, $w$ is unknown, and learning is the task of determining optimal parameter values from given data. The effect caused by random uncertainty, which is always present in the observation and creation of data, is denoted as noise $\epsilon$.

$$y = f(x; w) + \epsilon \tag{2}$$

$\epsilon$ is a random variable and represents the stirring up of regularity by random noise. Assuming that the regularity is a linear function of the parameters, the assumed regression model becomes the following equation.

$$f(x; \omega) = w\top h(x) = \sum_j w_j h_j(x) \tag{3}$$

This is linear regression. Model $f(x; w)$ fits the implicit function that represents data $D$. The mean squared error (MSE) is the error function that represents that measure.

$$E_D(w) = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}(x_n; w) - y_n)^2 \tag{4}$$

$\hat{y}(x_n; w)$ is the model prediction obtained by substituting the nth beauty power data into the model, where $y_n$ is the target value of the correct output that actually corresponds to the data. Thus $(\hat{y}(x_n; w) - y_n)^2$ is the difference between the prediction and the correct answer square. This is the error function, and the method of minimizing the mean squared error is the least squares method. Machine learning has a generalization error for arbitrary data, but this is approximated and minimized by the sample mean of the data. Assume that the data are given by equation (1) for the functional regularity f and the noise contribution. Suppose that $\epsilon$ follows a Gaussian distribution with mean 0 and variance $\sigma^2$. That is, if y itself is based on a Gaussian distribution with the estimated value $\hat{y}(x; w)$ as the mean, the entire model is expressed by the following equation.

$$\epsilon \sim N(\epsilon; 0, \sigma^2) \rightarrow y \sim P(\mathrm{y}|\mathrm{x} = x; w) = N(\mathrm{y}; \hat{y}(x; w), \sigma^2) \tag{5}$$

The maximum likelihood method can be used to bring this model closer to the distribution that generates the data.

$$L(w) \prod_n P(y_n|x_n; w) \tag{6}$$

Maximize the logarithm of the likelihood function above. In this model, $(x_n, y_n)$ is sampled

from a Gaussian distribution, so from the definition of Gaussian distribution, the log likelihood is the following equation.

$$\log \prod_n P(y_n|x_n;w) = -\frac{1}{2\sigma^2}\sum_n (\hat{y}(x_n;w) - y_n)^2 + const. \tag{7}$$

The maximization of this equation is the minimization of $(\hat{y}(x_n;w) - y_n)^2$, which is identical to the minimization of the mean squared error. From the above, the least squares method is the maximum likelihood method when the variation of the estimator from the functional model is assumed to follow a Gaussian distribution.

This study dealt with an engineering problem and data augmentation was used to compensate for the insufficient data. Uniform noises with a few percentages to original data value were added to the original data. Because the least squares method can be equated with the maximum likelihood method. Data augmentation was based on the distribution of the target variable, as shown in Equation (8).

$$n_i \times c_i = xn_{max} \tag{8}$$

$n_i$ is the number of data in the range to be augmented. $c_i$ is the individual augment fact . $x$ is the augmentation factor by definition. $n_{max}$ is the number of data in the mode-most data range.

## 2.2 Data and Research Methods

In this study, results of a collision simulation between a finite element method model of a car and a leg is predicted by machine learning.

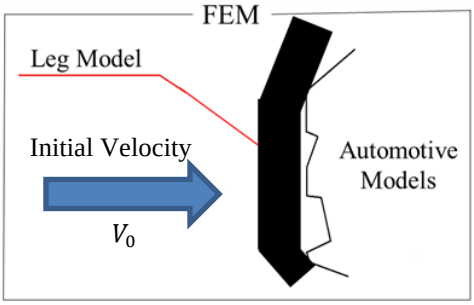


**Figure 1**: Simulation model using finite element method

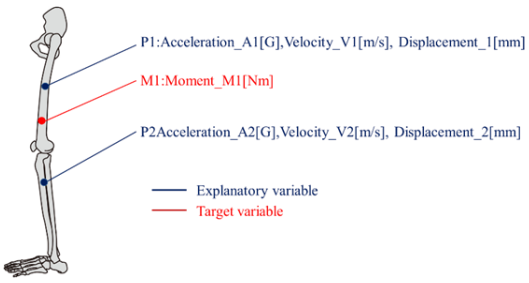


**Figure 2**: input-output point

The input values is acceleration, velocity, and displacement at points P1 and P2 and the value to be predicted is set the maximum bending moment at point M1. The distribution of the objective function is shown in figure 3.

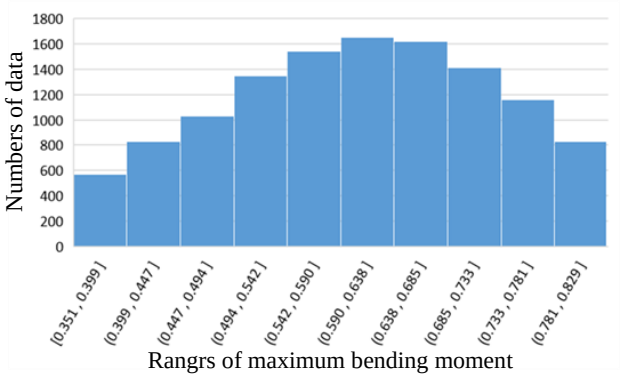


**Figure 3**: The maximum bending moment

Figure 4 shows a convolutional neural network in this study.

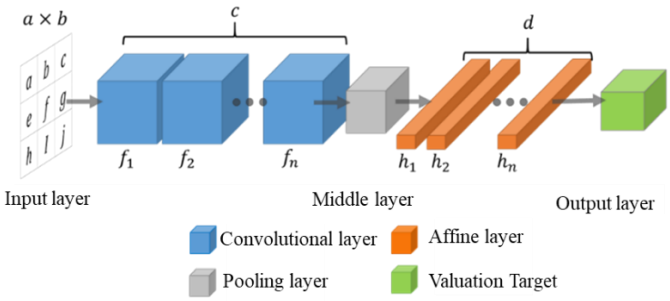


**Figure 4**: Convolutional neural network model

In this study, three convolutional layers and three all-coupled layers were set up. Input data consist of the acceleration, velocity, and displacement of the P1 and P2 points and their maximum values. All of data are regularly arranged and normalized as a $12 \times 12$ matrix.

## 3 RESULTS AND DISCUSSION

The number of the mode in Figure 3 is 1,654. According to Equation (8), the bin was augmented up to 16,540. The number is approximately 13.8 times greater than the original data. The augumented data sets is designated as case A.

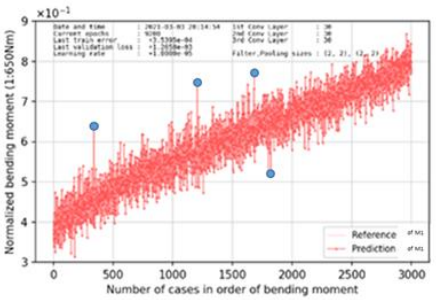


**Figure 5**: Prediction Results of Case A

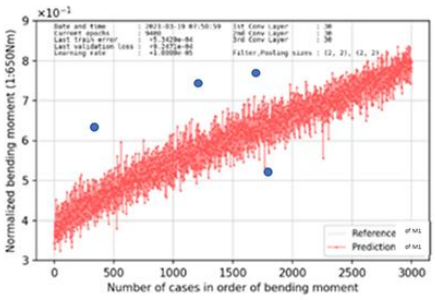


**Figure 6**: Prediction Results ofr Case B

The maximum error between the predicted and validated values is approximately 41.2%, and the last validation loss is $1.1838 \times 10^{-3}$. From this result, it can be seen that the predictions contain errors of 25% or more at the four points shown in Figure 5.

It is speculated that the large prediction error may be due to the small number of data at that point. Therefore, we augment the number of data consider to be insufficient by an additional factor of 1243 by a factor of 50, add them to the training data, and train them as Case B. After the expansion, the number of data is 19 times larger than the original. The maximum error is approximately 20.1%, and the last validation loss is $9.2471 \times 10^{-4}$.

Compared to Case A, extending the missing interval resulted in not only a decrease in the maximum error within that interval, but also a 17.3% increase in the overall accuracy of the learning in terms of last validation loss; the MSE decreased by 32.6%. A correlation diagram is shown in Figure 7.

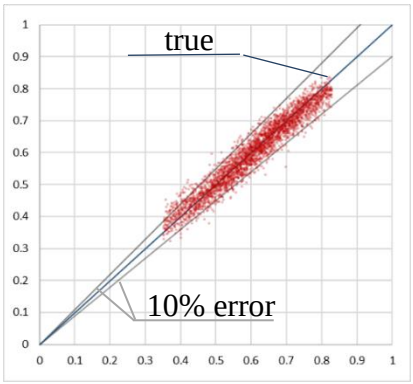


**Figure 7**: Regression line

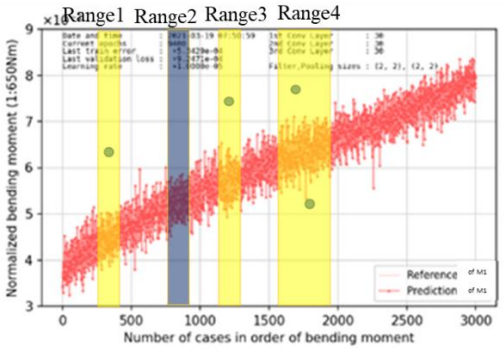


**Figure 8**: Range settings

If the insufficient data can be explored and estimated machine learning, it is expected to reduce the computational cost. We'd like to define the evaluation ranges as shown in Figure 8. Range 1 is for a bending moment range between 350 to 450 case numbers. The other ranges are also defined by the same manner as shown in Figure 8. For the estimation, 50 data including four points which is over 25% error randomly chosen from the groups shown in Figure 8. The data classified by the t-SNE method for displacement and maximum bending moment were classified by t-SNE. Range 2 does not include the results over 25% error.

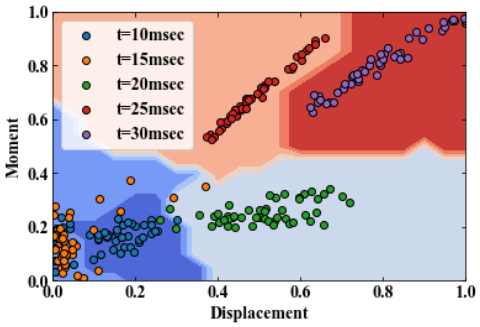


**Figure 9**: Results of t-SNE for range 1

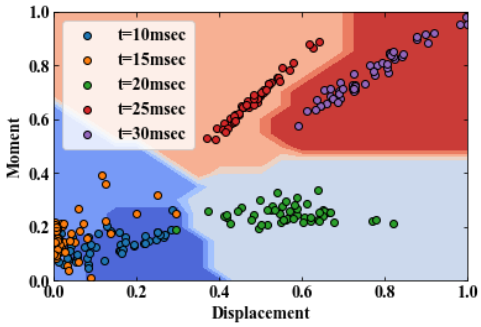


**Figure 10**: Results of t-SNE for range 2

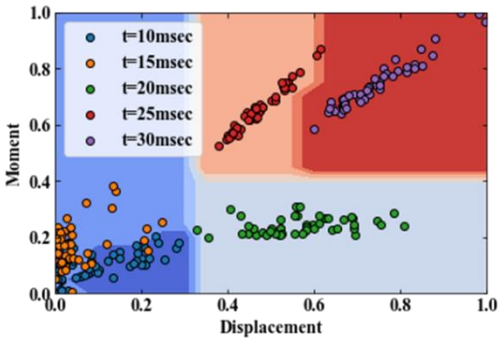


**Figure 11**: Results of t-SNE for range 3

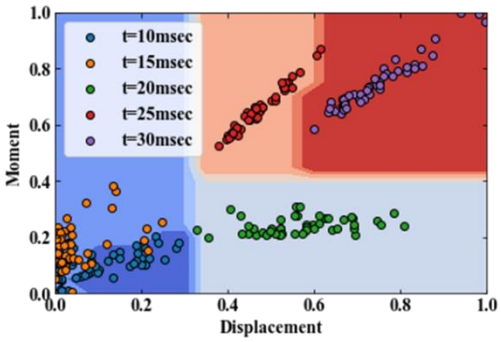


**Figure 12**: Results of t-SNE for range 4

Decision boundaries in t-SNE represent classification boundaries. This makes it possible to use linear interpolation, which is a data augmentation other than including noise.

The t-SNE allowed us to set and classify decision boundaries, however, we could not find any suggestions to get more accuracy from the boundaries. We claim that classification of parameters by time series is not effective, and that it is necessary to classify parameters according to their accuracy at characteristic times. In addition, when attempting linear interpolation using data of different classifications based on decision boundaries, it is necessary to verify whether the additional data to be interpolated can represent the implicit function for an engineering problem for evaluation of pedestrian protection.

## 4 CONCLUSION

The results of Case B show that data augmentation with low frequency of occurrence is effective in improving prediction accuracy. In order to use the t-SNE for data augmentation, the parameters should be classified by learning accuracy.

## REFERENCES

[1] Kentaro Yamamoto, Approach to reduce the computational cost of crash simulation using machine learning. *Proceedings of the conference on computational engineering and science*. (2021) **26**: B-05-03. (in Japanese)

[2] Kai Ogata, Construction What Surrogate Model Using Evaluation Safety Performance When Cars Clash Walking People. *Proceedings of the conference on computational engineering and science*. (2021) **26**: B-05-04. (in Japanese)

[3] Yoshitaka Wada, Application of convolutional neural network to regression problem for physical phenomenon. *Proceedings of the conference on computational engineering and science*. (2021) **26**: B-05-05. (in Japanese)