Scientific paper

Undersampling Strategy for Machine-learned Deterioration Regression Model in Concrete Bridges

Yuriko Okazaki1*, Shinichiro Okazaki2, Shingo Asamoto3 and Pang-jo Chun4

Received 27 May 2020, accepted 17 November 2020 doi:10.3151/jact.18.753

Abstract

Inspection data of actual concrete structures should be analyzed to elucidate the deterioration mechanism and construct a regression model. Although machine learning can be applied to this problem, inspection data are not suitable because machine learning targets big data with a uniform density and a balanced distribution. This study applies machine learning to a regression model of the crack damage grade in concrete bridges, using imbalanced inspection data. The model performance is improved by analyzing the influence of undersampling. Undersampling is conducted step-wise, and the models are constructed by learning all the undersampled data. The cross-validation of these models yielded the regression errors on each crack damage grade to evaluate the model performance considering the bias of data imbalance. Based on the results, the effect of undersampling on the model performance is analyzed, and the appropriate model is selected. Additionally, the influence of the model difference on the evaluation is investigated via historical change or factor analysis to confirm the effect of undersampling. This article not only presents a case study of a regression task for crack damage grades in concrete bridges, but also describes a strategy to maximize the use of imbalanced data for regression problems.

1. Introduction

Deterioration of concrete structures consists of not only physical damage from live loads but also material damage by factors such as carbonation and chloride penetration, which progress intricately affecting each other. Although various approaches, including experimentation and numerical calculation schemes (Maekawa et al. 2003; Tanaka et al. 2017), have been used to analyze the deterioration mechanism and predict the deterioration progress, the inspection data of actual concrete structures should be analyzed as they constitute a database that includes homogeneous information about on-site deterioration under different conditions.

Various techniques, such as the Markov chain model (Jiang et al. 1988; Ng and Moses 1998; Morcous 2006; Wellalage et al. 2013), multiple regression analysis (Veshosky et al. 1994), and Cox regression analysis (Fang et al. 2018) have been adopted to analyze these inspection results, extract the characteristics of the deterioration, and reflect them in predictions of deterioration progress. These techniques can capture the general deterioration trend of the target bridges, but not the trend of individual bridges.

Currently, owing to the recent rapid improvement in the processing speed of computers, machine learning is receiving renewed attention. Machine learning can extract intricately linked patterns of parameters hidden in large amounts of data, and thus, it has been used extensively in a broad range of applications (Kim and Kang 2010; Akagi et al. 2019). Although the use of machine learning is becoming common in concrete engineering, for example, in damage detection through structural health monitoring (Amezquita-Sanchez and Adeli 2015; Rafiei and Adeli 2017; Perez-Ramirez et al. 2019) and image recognition (Yamaguchi and Hashimoto 2010; Prasanna et al. 2014), its implementation in the assessment of durability and service life of infrastructure is still limited. The reason may be that inspection data of infrastructure are usually not dense, and they are obtained without equidistant intervals over time. These data are sparse because the data obtained by visual inspection requires time and infrastructure has a long life. Meanwhile, the concept of learning systems in machine learning usually assumes that the training sets are well balanced (Japkowicz 2000; Chawla et al. 2004), which is often not the case with inspection data. Inspection data usually indicate that the infrastructure is sound, and the frequency at which deterioration data is observed is rare; thus, an imbalance of data occurs. The lack of density, especially with the high dimensionality of explanatory variables, represents a challenge in imbalanced learning, which often causes models to fail at detecting rare patterns (Haixiang et al. 2017). Thus, this leads to difficulties in the application of machine learning to assess the durability of infrastructure.

The procedure to accumulate deterioration data, apply
machine learning to the inspection data of infrastructure, and analyze the deterioration mechanism is simple, but it requires time and involves a significant cost. Furthermore, it is irrelevant from an engineering perspective to not analyze the data and evaluate the deterioration mechanism until sufficient deterioration data has been accumulated. Therefore, our study aims to extract the deterioration mechanism as accurately as possible from existing inspection data, even if these data are still sparse and imbalanced.

In our previous study, crack damage in the main girders of concrete bridges was investigated (Okazaki et al. 2020). The applicability of machine learning to the regression problem of crack damage formation and propagation using inspection data was validated by comparing the performance of different learning algorithms and tuning hyper-parameters. This study focuses on the sampling of learning data as one of the data pre-processing processes and aims to construct a model that can extract hidden patterns in data efficiently and accurately by analyzing and interpreting the influence of undersampling on model performance.

Note that this article not only presents the case study of a regression task for crack damage grades in concrete main girders but also describes a strategy to maximize the use of imbalanced data for regression problems. We hope that the results provide significant insight to civil engineers who try to apply machine learning to data analysis.

The remainder of this paper is shown in Fig. 1. Section 2 presents the characteristics of the targeted data. Section 3 outlines the procedure of machine learning used in this study. Section 4 applies undersampling techniques to the learning data. Undersampling is conducted stepwise, and the regression models are constructed via machine learning of all undersampled data. Section 5 investigates the influence of the model difference on the evaluation by historical change or factor analysis to confirm the effect of undersampling. The conclusions are presented in Section 6.

2. Targeted data

The training data were collected from the existing inspection results of 1688 road bridges with concrete main girders in four prefectures of the Shikoku region in Japan. The data were recorded periodically over an 11-year period from 2005 to 2015. Figures 2 and 3 present the annual mean temperature and annual precipitation in the Shikoku region using a 1 km mesh, with the locations of the target bridges indicated by yellow circles. The temperature differences in Shikoku are minimal, whereas the precipitation differences are appreciable. There is little precipitation in Kagawa Prefecture, approximately 1000

---

Fig. 1 Flow diagram of the study.
mm, whereas Kochi Prefecture experiences the highest annual precipitation in any region in Japan, over 3000 mm. Most bridges are located near the coastlines of every prefecture of Shikoku. It is known that chloride ions and water leads to the deterioration of concrete structures; thus, these data of the Shikoku region are appropriate for analyzing the environmental influences on the bridges in Japan.

Figure 4 shows the span length histograms of the target bridges. For bridges with spans of different lengths, the most frequent value is considered. The target bridges include reinforced concrete (RC) and pre-stressed concrete (PC) bridges, and they are separately included in the histograms of Fig. 4. Approximately 80% of RC bridges are less than 7.5 m in span length, whereas 90% of PC bridges are greater than 7.5 m and vary considerably, up to a few dozen meters. Figure 5 is a histogram of the years in service of the target bridges at the time of inspection. It is observed that the majority of bridges have been in service for less than 50 years.

The inspection data includes the type of damage and the grade assigned for the different elements, obtained by dividing the member of the target bridge into intervals (MLIT 2018b), as shown in Fig. 6. The left side of Table 1 presents 14 types of damages; this study aims to construct a regression model of the crack damage grade of

| Damage type | Damage grade | Max. Crack Width \((C_{\text{max}})\), (mm) | Min. Crack Interval \((C_{\text{min}})\), (m) |
|-------------|--------------|----------------------------------|------------------|
| Cracked     | Repair/repair damage | - | - |
| Deck cracked | Lifting | Abnormal noise | Abnormal noise or vibration |
| Peeling or rebar exposure | Gap anomaly | Abnormal deflection |
| Leakage or free lime | Bad anchorage zone | Deformation | fracture |
| Loose knot | Color change or deterioration | | |
| | | | |

Table 1 Damage types and grades used for periodic inspections.

![Figure 2](image2.png)

Fig. 2 Target bridges and annual mean temperature in the Shikoku region.

![Figure 3](image3.png)

Fig. 3 Target bridges and annual precipitation in the Shikoku region.

![Figure 4](image4.png)

Fig. 4 Span length of target bridges.

![Figure 5](image5.png)

Fig. 5 Span length of target bridges.
the main girders as it is the most frequently observed damage. The right side of Table 1 lists the damage grades, which extend from “a” (sound) to “e” (the most damaged). The definitions of grade for other damage types are described in the procedure (MLIT 2018a). The letter grades of damage correspond to discrete values, which is convenient for ranking purposes. However, because the actual progress of damage is not discontinuous and discrete but is rather continuous and gradual, the letter grades are replaced with numerical values from five (sound) to one (the most damaged) to form the machine learning dataset. For each target bridge, periodic inspection is conducted every five years (MLIT 2018a); in addition, there are only one to four inspection results because a significant amount of time has not passed since the transition to the current inspection guidelines. Thus, as shown in Fig. 7, some bridges lack inspection data for a period after the start of service.

Furthermore, because little time has passed since the transition to the current inspection guidelines, there are only one to four inspection results for each target bridge. Hence, there are few target bridges in service for the same number of years at the time of inspection, and it is difficult to compare the damage grade even between bridges of the same service age. Such a lack of data is known to reduce the accuracy of the model. The data are thus interpolated over time as much as possible according to the following rules:

First, the initial damage grade of year zero is set to 5, assuming that the bridges were sound at the start of their service. Subsequently, to simplify the alignment of the time axes of the target bridges, the years in service of the training data are rounded to intervals of 5 years. When there is a lack of available data and the grade of the data prior to the unavailable period is the same as that following the unavailable period, the missing data are interpolated as being of the same grade, assuming that the damage grade did not change during the intervening period. Furthermore, data showing a recovery of the grade compared with the previous inspection are eliminated because of the possibility of repair. Figure 7 shows an example of the interpolation of actual data on the time axis according to the aforementioned rules. The number of datasets after interpolation is 262,569.

Figure 8 shows histograms of the crack damage grade in the training data divided into RC and PC. The vertical axis indicates the frequency of data for each type of bridge, and the numeric characters on the bars show the number of data. In total, 98% of the data for PC bridges is of grade 5 and (i.e., sound). For RC bridges, most of the data, 87%, are also of grade 5; the corresponding percentage of data for other grades is 1% to 6%. The inspection results of RC bridges indicate more deterioration than PC bridges, and the difference is because PC is strengthened by tensioning during construction to prevent cracking. Additionally, most of the bridges built during the period of high economic growth were RC; PC bridges are relatively recent. The target data include other factors for crack propagation in concrete structures, and the following sections will construct the regression model and conduct a factor analysis to analyze their detailed influence.
3. Machine learning

3.1 Procedure of machine learning

The procedure of machine learning involves three main steps: data pre-processing, feature selection to determine the most significant set of explanatory variables for prediction, and data training to construct the regression model. Data pre-processing is the most important phase of machine learning. Raw data often include out-of-range and missing values. Analysis of data with these values can produce misleading results. Thus, data pre-processing should be conducted prior to data training. It includes cleaning to delete abnormal and improbable data, interpolation for missing data, and normalization to adjust explanatory values measured on different scales to a notionally common scale. Additionally, data sampling may help to improve the model performance when the training data is imbalanced.

The imbalance problem typically occurs in classification problems where the target variable is nominal, when there are many more instances of some classes than of others (Chawla et al. 2004; Haixiang et al. 2017). In such cases, the standard classifiers tend to be overwhelmed by large classes and ignore the small ones. The problem can be seen in a regression task where the target value is numeric, when there are few examples with rare target values (Torgo et al. 2015). Various solutions to this problem have been previously proposed at the algorithmic (Galar et al. 2013; Thanathamthee and Lursinsap 2013; Maalouf and Siddiqi 2014), evaluation metrics (Sun et al. 2007), and data levels (Garcia et al. 2012; Dubey et al. 2014). At the data level, the solutions include many different forms of resampling, such as random oversampling, random undersampling, directed undersampling where the choice of examples to eliminate is informed, for example, by the nearest neighbor algorithm, the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002), and combinations of the above techniques. The differences between these techniques have been reported in previous research (Zhou 2013; Loyola-González et al. 2016); the present study focuses on undersampling. Undersampling was the earliest prevailing technique used to alleviate the imbalance in the dataset. It removes samples from the majority class randomly or with consideration of data similarity. However, it can potentially remove useful or important samples. The quantity to undersample is usually detected empirically, and there is no clear metric or strategy. Imbalanced data represents a problem when it is adopted in both classification tasks, where the target variable is nominal, and regression tasks, where it is continuous; some studies have already addressed this (Torgo et al. 2015; Torgo 2016). They focus on the continuous response values and improve the imbalance of the dataset by considering their similarity and significance. Explanatory variables also have this problem; that is, how to determine the similarity and which value should be removed from the dataset as not being independent but a duplicated one. In other words, what the significant digits are of each explanatory variable. A detailed study on this problem has not been previously reported.

This study provides an undersampling strategy for data with explanatory variables with continuous values to construct an effective and accurate regression model from existing data. The targeted data is imbalanced as shown in Fig. 8, in which the majority of data is grade 5, that is, sound data, with the minority of data being under grade 4, that is, deterioration data. The data are undersampled stepwise, and each data point is learned to construct regression models. The effect of each undersampling on the model performance is investigated by comparing these models. First, the overlapped data for all explanatory values and the response value are simply deleted from the dataset, which was discussed in Section 2. Furthermore, the data of sound grade 5, which is the majority of the learning data, is undersampled based on the nearest neighbor algorithm in a stepwise manner. The details of undersampling with the nearest neighbor algorithm are presented in Section 4.1. Each data point is learned to construct the regression models, and these model performances are compared. The result leads to the selection of the appropriate undersampled data to construct the model more effectively and accurately than the original models. Finally, the influence of the difference in learning data obtained by undersampling is investigated in detail. It is noted that feature selection, that is, the selection of explanatory variables, is conducted prior to the undersampling. The details of this are provided in our previous paper (Okazaki et al. 2020) and are not included here. The explanatory variables by feature selection are listed in Table 2.

3.2 Machine learning algorithm

Machine learning is a technique that allows computers to mimic human learning abilities. It can recognize patterns and structures hidden in large amounts of data without assuming a predetermined equation as a model. Many machine learning algorithms have been proposed and are freely available. The present study adopts the Gaussian process regression (GPR) (Rasmussen 2006) as the learning algorithm, as it can fit the trend even for a local characteristic of the target data based on a performance comparison of the models constructed with different common algorithms, namely, multiple linear regression, support vector machine, decision trees, artificial neural network, and GPR (Okazaki et al. 2020).

The GPR model, which is a Bayesian nonparametric kernel-based probabilistic model, is outlined here (Bishop 2006). Consider a model defined in terms of a linear combination of fixed basis functions given by the elements of the vector \( \phi(x) \) such that,

\[
y(x) = w^T \phi(x)
\]

where \( x \) is the input vector and \( w \) is the M-dimensional
The application of Gaussian process models to the problem of regression requires consideration of noise in the observed target values as,

\[ t_n = y_n + \varepsilon_n \quad (n = 1 - N) \]

where \( y_n = y(x_n) \), and \( \varepsilon \) is a random noise variable whose value is chosen independently for each observation \( n \). Considering that \( w \) and \( \varepsilon \) follow Gaussian distributions, the marginal distribution of \( t \) is given by in Eq. (3).

\[ p(t) = N(t \mid 0, C) \]  

An important result is that \( p(t) \) is obtained from only the mean and covariance matrix \( C \) has the elements shown in Eq. (4).

\[ C(x_n, x_m) = k(x_n, x_m) + \beta^{-1} \delta_{nm} \]

where \( k \) is a kernel function, \( \beta \) is a hyper-parameter indicating the variance of \( \varepsilon \), and \( \delta \) is the Kronecker delta. The kernel function determines how the response at a point \( x_n \) is affected by the responses at other points \( x_m \). For many standard kernel functions, such as the squared exponential kernel, exponential kernel, Matern 3/2 kernel, and rational quadratic kernel, the kernel parameters are based on the signal standard deviation \( \sigma_f \) and the characteristic length scale \( \sigma_l \). The present study adopts the rational quadratic kernel expressed in Eq. (5) because a model incorporating this kernel provides a better result of cross validation than models incorporating the other kernel functions mentioned earlier when using the data targeted in this study. The rational quadratic kernel is,

\[ k(x_n, x_m / \theta) = \sigma_f^2 (1 + \frac{r^2}{2\alpha \sigma_l^2} + \frac{\alpha^2}{2}) \]

where \( \alpha \) is a positive-valued scale-mixture parameter, and \( r = \sqrt{(x_i - x_j)^T(x_i - x_j)} \) is the Euclidean distance between \( x_i \) and \( x_j \). The characteristic length scale \( \sigma_l \) briefly defines how far apart the input values \( x_i \) can be for the response values to become uncorrelated. Both \( \sigma_f \) and \( \sigma_l \) need to be greater than zero, and this can be enforced by the unconstrained parameterization vector \( \theta \), such that \( \theta_l = \log \sigma_l \) and \( \theta_f = \log \sigma_f \).

The parameters in Eqs. (4) and (5) are estimated from the data by maximizing a log likelihood function expressed in Eq. (6) while training the GPR model:

\[ \ln p(t \mid \theta) = \ln[C_N] - \frac{1}{2} t^T C_N^{-1} t - \frac{N}{2} \ln(2\pi) \]  

For \( N \) training of data, the prediction of the objective variable \( t_{N+1} \) for a new input vector \( x_{N+1} \) is equivalent to finding the conditional distribution \( p(t_{N+1} \mid t) \). Using vector \( k \) with element \( k(x_n, x_{N+1}) \) and scalar \( c = k(x_{N+1}, x_{N+1}) + \beta \), the mean and covariance for \( p(t_{N+1} \mid t) \) are derived as,

\[ m(x_{N+1}) = k^T C_N^{-1} t \]  
\[ \sigma^2(x_{N+1}) = c - k^T C_N^{-1} k \]

GPR is a lazy learning algorithm, in which the generalization of the training data is delayed until a query is made to the system, as opposed to other learning algorithms, where the system tries to generalize the training data before receiving queries. The GPR model measures the similarity between points to predict the value for an unseen point from the training data; it has an arithmetic learning cost of \( O(n^3) \), where \( n \) is the size of the training dataset. Therefore, adequate data undersampling for GPR models leads not only to the improvement in model performance by alleviating the data imbalance, but also to the reduction of the arithmetic cost, which has engineering significance.

---

Table 2: Explanatory variables based on feature selection.

| Category                        | Item                                      |
|---------------------------------|-------------------------------------------|
| Bridge specifications           | Bridge length                             |
| Design conditions               | Span length                               |
| Load conditions                 | Total width                               |
| Environmental condition         | Bridge type                               |
| Time axis                       | Total number of spans                     |
| Span variability                | Cast location                             |
|                                 |                                           |
|                                 | Items selected as explanatory variables   |
4. Undersampling and model construction using data learning

4.1 Undersampling of dataset
As stated in Section 3.1, the overlapped data was first deleted from 262,569 data points by eliminating the data with the same target and explanatory values, and it means that duplicated data in each bridge was eliminated. Subsequently, the sound grade 5 data were undersampled in a stepwise manner based on the nearest neighbor algorithm (Friedman et al. 1977), and datasets at each undersampling step were obtained. Figure 9 shows the detailed algorithm of the undersampling in this study. For an individual sample of grade 5 data, similar data are extracted by the nearest neighbor method; whether it is removed or not from the dataset is determined by the following rules:

(1) When the majority of the neighbor samples have different response values, that is, grades 1 to 4, the targeted sample of grade 5 is eliminated as the anomaly.

(2) When all the neighboring samples have the same response value, that is, grade 5, the targeted sample is eliminated as the region is too congested. Additionally, the nearest neighbor sample is precluded from being an elimination candidate as the neighbor of other individuals.

It is noted that the explainable values were normalized before undersampling to compare the similarity between samples by the same scale.

The 3-nearest neighbor algorithm was used in this study, and undersampling by the algorithm shown in Fig. 9 was repeated eight times until there was no eliminated data. It followed that 10 datasets were obtained, which included original data, overlapping deleted data, and 8 datasets from the undersampling based on the nearest neighbor method. Figure 10 indicates the number of samples in the 10 datasets. The original data with 262,569 samples was reduced to 21,155 by overlapping deletion, and additionally, the number of data was reduced to between 11,166 to 5,188 by undersampling. The relative histograms of crack damage ranks for these datasets are shown in Fig. 11. The proportion of grade 5 data to the original dataset is 97%, that to the overlapping deleted dataset is 86%, and that to datasets by undersampling based on the nearest neighbor method is reduced to 57% to 41%, respectively, which indicates the alleviation of data imbalance, although it is not ideal.

4.2 Model construction by data learning
This section constructs 10 regression models by learning the 10 datasets obtained as described in Section 4.1. Machine learning often learns the errors related to outliers; the generalization ability is thus reduced in the so-called “overfitting.” Therefore, the resubstitution error, which is the difference between the response value of the learning data and the predictions, cannot evaluate the generalization ability of the model, and having a low resubstitution error does not guarantee good predictions for new data. To evaluate the generalization ability of the...
model constructed by machine learning, cross-validation, leave-one-out, and hold-out techniques are widely used. These techniques divide the available dataset into “training dataset” and “test dataset.” A model learns the training dataset; the test dataset, which is the first observed data for the model, is used to validate the model performance. The techniques are different in the way they divide the dataset; this study adopted the 50-fold cross validation in order to evaluate the detailed generalization ability of the models considering the accuracy of evaluation and the computational cost. In 50-fold cross validation, the data was partitioned into 50 groups; 50 rounds of validation were performed using different partitions, and the validation results were combined over the 50 rounds to reduce variability. Figure 12 indicates the concept and procedure of the cross-validation in this study. The inspection data were obtained according to each bridge, not randomly, and thus, partitioning was performed at the bridge level, not the data level. A total of 1688 bridges were partitioned into 50 groups; the data of 49 groups were learned, and the data of the remaining one group were used for validation. The root-mean-square-deviation (RMSE) expressed in Eq. (9) and mean absolute error (MAE) in Eq. (10) were adopted as the metrics of the evaluation.

\[
\text{RMSE} = \sqrt{\frac{\sum (y_{\text{obs},i} - y_{\text{pred},i})^2}{n}} \tag{9}
\]

\[
\text{MAE} = \frac{\sum |y_{\text{obs},i} - y_{\text{pred},i}|}{n} \tag{10}
\]

where \(y_{\text{obs},i}\) is the observed value of \(i\)-th sample, \(y_{\text{pred},i}\) is the prediction of \(i\)-th sample, and \(n\) is the number of samples in the dataset.

The MAE is the average of the absolute values of errors between the observed values and the predictions and is fundamentally easy to understand. The RMSE is the square root of the second sample moment of the differences between observed values and predictions, and it has a high sensitivity to the errors for the outliers.

It is noted that some hyper-parameters shown in Eq. (5) should be designed for the GPR model learning. In this study, they were optimized to control the progression of the crack damage grade by the regression model more consistently with the knowledge of concrete engineering, while maintaining a low RMSE for the test data (Okazaki et al. 2020).

For classification tasks in imbalanced learning, the evaluation metrics that are not to be biased toward the majority class are often utilized; examples of these metrics include the AUC, G-mean, and F-measure (Thanathanthatee and Lursinsap 2013; Haixiang et al. 2017). The present study adopts RMSE and MAE on each crack damage grade as the metric for evaluating the model performance considering the bias of data imbalance. Figure 13 shows the results of the 50-fold cross-validation for 10 regression models, RMSE, and MAE. The horizontal axis of these graphs indicates the number of samples in the datasets, and the vertical axis represents the RMSE and MAE in the top and bottom graphs, respectively. The RMSE and MAE were calculated for the test data with all grades, and for each grade. The results for all grades show that there is no clear
difference in RMSE between the full data model from 262,569 samples, the overlapping deleted model from 21,155 samples, the undersampling (1) model from 11,166 samples, and the undersampling (2) model from 7,133 samples. For these models, the MAE shows only a slightly increasing tendency. In contrast, the models from the data obtained after undersampling (2) show a considerable increase in both RMSE and MAE, which means that the model performance degrades despite the lower reduction in the number of samples in the data compared to the before undersampling (2).

For the results of each grade, both RMSE and MAE indicate almost the same trend. The RMSE and MAE for grade 5 increase, that is, the model performance decreases as the number of samples in the data decreases. The increasing trend is not obvious until undersampling (2) and becomes considerable after it is the same as the result for all grades. For grades 1 to 3, they decrease, that is, the model performance improves until undersampling (2), as the number of samples decreases, and there is no obvious change after undersampling (2). From these results, it is evident that the data reduction until undersampling (2) could improve the model performance for the data of low grade, while maintaining the performance for the data of grade 5, which results in the maintenance of RMSE and MAE for the test data with all grades. The data reduction after undersampling (2) considerably decreases the model performance for grade 5 data, while there is no influence on the low grade.

It is noted that the aforementioned trend of the data for grade 5 and that for all the grades is due to the distribution of test data. As shown in Fig. 11, the data still consist mostly of grade 5 even after undersampling. Therefore, the regression errors for the data of grade 5 significantly affect the RMSE and MAE for all the data. From the above, it is important to investigate how the model can track the data considering its distribution to evaluate and interpret accurately the performance of the model constructed by machine learning.

Figure 14 shows the relation between the number of samples in the data and the learning time. The results were run on a computer with a dual 16-core Intel Xeon processor running at 2.10 GHz using 128 GB of RAM. While the learning of full data with 262,569 samples requires approximately 30,000 s, that is, 8 h, the overlapping deleted data with 21,155 samples and the undersampling (2) data with 7,133 samples took only 15 s and 1 s, respectively. As stated earlier, the undersampling (2) model learned the data with samples having just one thirty-sixth of the full data and has the same generalization ability as the full data model. Thus, this undersampling can considerably reduce the computational cost without model performance degradation.

The data obtained from undersampling after undersampling (2) lead to the degradation of regression performance, especially for the data of grade 5, although there is a small improvement in the data of low grades. It is obviously necessary to ensure the regression performance not only for the deterioration data of low grade but also for the sound data of grade 5 to accurately track the progress of the crack damage grade of concrete structures. Thus, this study selected the model that learned the undersampling (2) data as the appropriate one.

The undersampling after undersampling (2) improve the model performance, which leads to the interpretation that the data removed at the steps from overlapping deletion to undersampling (2) is the so-called “duplicated” data; these data are unnecessary to construct the model and should be removed. Undersampling after undersampling (2) removes important examples which are “necessary” data to construct the model. By investigating the interfaces between the “duplicated” and “necessary” data, undersampling in machine learning for the present regression task is optimized.

To analyze the subject, the differences between the removed data and their nearest data were investigated by dividing the data-removal steps into two groups: (A) group from overlapping deletion to undersampling (2), in which the removed data is “duplicated”, and (B) group from undersampling (3) to undersampling (8), in which the removed data is “necessary.” Figure 15 shows the differences between the removed data and their nearest data for the main explanatory variables, that is, span length, completed years, minimum distance from coastline, and annual precipitation. It is noted that this investigation aims to determine the general influence of each explanatory variable, although it is impossible to grasp the influence because undersampling by the nearest neighbor method treats the combined similarity of all explanatory variables as the metric to elect the neighbor. For span length, the differences between the removed data and the nearest data for group (A) are almost always less than a few meters, while those over 3 m are the majority for group (B). This indicates that the differences over 3 m should not be removed from the training data, and are significant figures for span length. In the same way, differences of approximately 5 years for completed years, a few hundred meters for minimum distance from coastline, and 50 mm/year for annual precipitation should not be removed.

The investigation provides an understanding of the
significant figures required as explanatory variables to the target value; it also provides instructive information for further data analysis or data acquisition in the future. There is a possibility that the significant figures may differ according to the range of explanatory variables, and this will be investigated in a future work.

5. Detailed investigation of influence of model difference

5.1 Comparison between the observed values and predictions

The previous section evaluated the generalization abilities of each model using a 50-fold cross validation. This section focuses on three models based on full data with 262,569 samples, overlapping deletion data with 21,155 samples, and undersampling (2) data with 7,133 samples. The comparison between the observed data and the predictions of these models was performed to confirm the effect of undersampling.

Figure 16 shows the histograms of predictions of the three models for each grade of observed values. The predictions were rounded to the nearest integer to allow easy comparison of the observed values. For the data of grade 5, most predictions of all models also indicate the same grade. For the observed value of grade 4, the model from overlapping deletion data, which is shown with striped bars, can track them more accurately. The accuracy of regression for the data of grades 1 to 3 improves with a decrease in the number of samples in learning data, and it is the highest in the model from undersampling (2) data, which is shown with yellow bars. These trends are consistent with the results of the 50-fold cross validation in the previous section.

The mode of regression values for grade 1 is grade 2, which does not coincide with the observed value, and the accuracy is relatively low compared with that of the other grades. This seems to account for the minority of grade 1 data even after undersampling. It may also indicate that the data for grade 1 need another explanatory variable.
not included in Table 2, or that the number of data is not sufficient to construct an accurate model. The addition of available data in these regions to the training data or coupling the data with a parametric model-driven approach is required to improve the performance of the model in future work.

5.2 Historical change comparison
The previous section investigated the model performance by resubstituting errors to the learning data. The regression problem for the crack damage grade of concrete bridges targeted in this study requires the accurate tracking performance of the historical change for each bridge, including not only the range of inspection results obtained but also that without them. Figure 17 shows an example of the historical change comparison between the inspection results and predictions from the three models. The bridge of the example had inspection data after it had been in service for 30 and 35 years, and these inspection results are shown in Fig. 18 as a color map of the crack damage grade for each element. The elements without color, Nos. 101, 102, 103, and 601, are without “presence of water leakage,” which is one of the explanatory values, while the other colored elements are with it. It is known that the “presence of water leakage” affects the deterioration of concrete structures; thus, this section discusses the influence of three different models on the historical change in the crack damage grade without the presence of water leakage, to simplify interpretation of the comparison. The inspection results provide different values even when they are of the same bridge. At inspections after 30 and 35 years, the majority of elements provide a sound result, grade 5, and the deterioration progress is confirmed in the minority. Figure 17 shows predictions of the three models with the historical change, which starts to deteriorate at 30 years of service. The predictions at this time are grade 4 for the model from the full data with 262,569 samples, grade 3 for that from the overlapping deletion data with 21,155 samples, and grade 2 for that from undersampling (2) data with 7133 samples. The predictions degrade as the number of samples in the learning data decreases. This is because the model based on full data learned the inspection data of all elements, and thus, it tracked the average grade of crack damage for the bridge. The models based on overlapping deletion data and undersampling (2) data learned the inspection data in which the grade 5 sound data may be partly removed, and thus, they have a trend to track the crack damage of more deteriorated elements.

It is noted that the predictions of the 3 models after 40 years of service indicate the recovery of crack damage grade. The recovery is not consistent with the concrete engineering; that is, the crack damage grade of concrete structures worsens or remains unchanged as time proceeds, and there is no recovery without repair. The recovery of the grade compared with the previous inspection has been eliminated in this study, and the behavior of regression values is thus due to the high sparseness of data beyond 50 years. The example bridge is intended for service during up to 35 years, and the regression value after 35 years is therefore significantly affected by the inspection results of other bridges. The inspections for bridges that have served for more than 50 years are especially sparse. They remain at grade 5 because most bridges have already degraded and been demolished after 50 years, and only the sound bridges remain in the region. The regression performance of the region will be improved by accumulating data or coupling with a parametric model-driven approach in a future task, in a similar way to the data of grade 1 stated in Section 5.2.

5.3 Influence of environmental factors on crack damage grade
The deterioration of a concrete structure is strongly affected by the environmental conditions, which leads to the complexity of the deterioration mechanism. Therefore, this section presents an analysis of the influence of environmental factors on crack damage grade using the three models, and compares these results to understand the difference.

“Minimum distance from coastline” and “annual precipitation” are the factors taken as representative environmental factors that may influence the deterioration of concrete structures, and the relations based on regression values between them in the crack damage grade are shown in Fig. 19 for each period of 10 years in service.
Note that these results are calculated for the most frequently type of structure in the target bridges, that is, an RC bridge with a 10 m span length. The other parameters are shown at the bottom of Fig. 19. There is no clear difference between the results from the full data model and those from the overlapping deletion data model. At 10 years of service, the results with any distance from the coastline and annual precipitation indicate grade 5, no deterioration, and the grade of the region over 10 000 m from the coastline and high precipitation degrades to 4 at 20 years in service. Furthermore, at 30 years, the grade of the region with distance from coastline of 10 000 to 15 000 m and annual precipitation of 2200 to 2700 mm degrades to 3, and the grade 4 area expands to the region close to the coastline with 2200 mm of precipitation. At 40 years, the region again degrades to grade 3. Thus, the models from full data and overlapping deletion data have the same sensitivity for the case shown in this figure. Meanwhile, the model based on sampling (2) data provides a slightly different result, for example, at 40 years. It appears to capture the deterioration risk more severely. The results of the detailed investigations in this section explain the effect of undersampling on model performance, and demonstrate that the model with appropriate
undersampling of learning data can capture the deterioration risk more severely.

6. Conclusions

Machine learning was employed to develop a regression model for the crack damage grade in the main girders of concrete bridges by using imbalanced inspection data. The model performance was improved by analyzing and interpreting the influence of undersampling.

1) Undersampling was conducted stepwise and the regression models were constructed by learning each undersampled data. The result of comparing their RMSE and MAE on 50-hold cross validations led to the selection of the appropriate model. This can considerably reduce the computational cost.

2) The RMSE and MAE were calculated for not only the full test data but also the data of each crack damage grade. When the data that should be evaluated accurately is a minority, it is important to investigate how the model can track the data considering the distribution with the aim to evaluate and interpret the model performance accurately.

3) The RMSE and MAE were used to judge the interfaces between “duplicated” data, which should be removed, and “necessary” data. This provided an understanding of the significant figures required as the explanatory variables to the target value.

4) The influence of model difference on the evaluation according to the historical change and environmental factors was investigated to confirm the effect of undersampling. The results demonstrate that the model developed by appropriate undersampling of learning data can capture the deterioration risk more severely. This study focused on the undersampling technique to alleviate data imbalance. More effective resampling may be possible, by examining the efficacy of different types or combinations of sampling (Dubey et al. 2014), and it will be studied in the future.

Acknowledgements

This work was supported by Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Construction Technology Research and Development Grant Program (CRD), the Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP) and JSPS KAKENHI Grant Number 17H04932.

References

Akagi, M., Nakamura, Y., Higaki, T., Narita, K., Honda, Y., Zhou, J., Yu, Z., Akino, N. and Awai, K., (2019). “Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT.” European Radiology, 29(11), 6163-6171.

Amezquita-Sanchez, J. P. and Adeli, H., (2015). “Synchrosqueezed wavelet transform-fractality model for locating, detecting, and quantifying damage in smart highrise building structures.” Smart Materials and Structures, 24(6), Article ID 065034.

Bishop, C. M., (2006). “Pattern recognition and machine learning.” New York: Springer Science and Business Media LLC.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P., (2002). “SMOTE: Synthetic minority over-sampling technique.” Journal of Artificial Intelligence Research, 16, 321-357.

Chawla, N. V., Japkowicz, N. and Kotcz, A., (2004). “Editorial: Special issue on learning from imbalanced data sets.” ACM Special Interest Group on Knowledge Discovery in Data Explorations Newsletter, 6(1), 1-6.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J. and Initiative, A. S. D. N., (2014). “Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study.” NeuroImage, 87, 220-241.

Fang, J., Ishida, T. and Yamazaki, T., (2018). “Quantitative evaluation of risk factors affecting the deterioration of RC deck slab components in East Japan and Tokyo regions using survival analysis.” Applied Sciences, 8(9), 1470.

Friedman, J. H., Bentley, J. L. and Finkel, R. A., (1977). “An algorithm for finding best matches in logarithmic expected time.” ACM Transactions on Mathematical Software, 3(3), 209-226.

Galar, M., Fernández, A., Barrenechea, E. and Herrera, F., (2013). “EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling.” Pattern Recognition, 46(12), 3460-3471.

García, V., Sánchez, J. S. and Mollineda, R. A., (2012). “On the effectiveness of preprocessing methods when dealing with different levels of class imbalance.” Knowledge-Based Systems, 25(1), 13-21.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G. and Bing, G., (2017). “Learning from class-imbalanced data: Review of methods and applications.” Expert Systems with Applications, 73, 229-239.

Japkowicz, N., (2000). “Learning from imbalanced data sets: A comparison of various strategies.” In: Proc. AAAI Workshop on Learning from Imbalanced Data Sets, Austin, Texas, USA 31 July 2000. Menlo Park, California, USA: The AAAI Press, 10-15.

Jiang, Y., Saito, M. and Sinha, K. C., (1988). “Bridge performance prediction model using the Markov chain.” Transportation Research Record, 1180(1), 25-32.

Kim, M. J. and Kang, D. K., (2010). “Ensemble with neural networks for bankruptcy prediction.” Expert Systems with Applications, 37(4), 3373-3379.

Loyoa-González, O., Martinez-Trinidad, J. F., Carrasco-Ochoa, J. A. and García-Borroto, M., (2016). “Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases.” Neurocomputing, 175, 935-947.

Maalouf, M. and Siddiqi, M., (2014). “Weighted logistic regression for large-scale imbalanced and rare events data.” Knowledge-Based Systems, 59, 142-148.
Maekawa, K., Ishida, T. and Kishi, T., (2003). “Multi-scale modeling of concrete performance.” *Journal of Advanced Concrete Technology*, 1(2), 91-126.

MLIT, (2018a). “Bridge periodic inspection procedures.” Tokyo: Road and Disaster Prevention Section of the Ministry of Land, Infrastructure, Transport and Tourism. (in Japanese)

MLIT, (2018b). “Road bridge periodic inspection procedures.” Tokyo: Road and Disaster Prevention Section of the Ministry of Land, Infrastructure, Transport and Tourism. (in Japanese)

Morcous, G., (2006). “Performance prediction of bridge deck systems using Markov chains.” *Journal of Performance of Constructed Facilities*, 20(2), 146-155.

Ng, S. K. and Moses, F., (1998). “Bridge deterioration modeling using semi-Markov theory.” In: N. Shiraishi, M. Shinozuka and Y. K. Wen, Eds. *Proc. 7th International Conference on Structural Safety and Reliability*, Kyoto, Japan 24-28 November 1997. Amsterdam: A. A. Balkema Publishers, 1, 113-120.

Okazaki, Y., Okazaki, S., Asamoto, S. and Chun, P. J., (2020). “Applicability of machine learning to a crack model in concrete bridges.” *Computer Aided Civil and Infrastructure Engineering*, 35, 775-792.

Perez-Ramirez, C. A., Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., Adeli, H., Dominguez-Gonzalez, A. and Romero-Troncoso, R. J., (2019). “Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings.” *Engineering Structures*, 178, 603-615.

Prasanna, P., Dana, K. J., Gucunski, N., Basily, B. B., La, H. M., Lim, R. S. and Parvardeh, H., (2014). “Automated crack detection on concrete bridges.” *IEEE Transactions on Automation Science and Engineering*, 13(2), 591-599.

Rafiei, M. H. and Adeli, H., (2017). “A novel machine learning-based algorithm to detect damage in high-rise building structures.” *The Structural Design of Tall and Special Buildings*, 26(18), e1400.

Rasmussen, C. E., (2006). “Gaussian processes in machine learning.” In: O. Bousquet, U. von Luxburg and G. Rätsch, Eds. *Summer School on Machine Learning, ML 2003: Advanced Lectures on Machine Learning, Lecture Notes in Computer Science*. Berlin and Heidelberg: Springer-Verlag, 63-71.

Sun, Y., Kamel, M. S., Wong, A. K. and Wang, Y., (2007). “Cost-sensitive boosting for classification of imbalanced data.” *Pattern Recognition*, 40(12), 3358-3378.

Tanaka, Y., Takahashi, Y. and Maekawa, K., (2017). “Remaining fatigue life assessment of damaged RC decks: Data assimilation of multi-scale model and site inspection.” *Journal of Advanced Concrete Technology*, 15(7), 328-345.

Thanathamthee, P. and Lursinsap, C., (2013). “Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques.” *Pattern Recognition Letters*, 34(12), 1339-1347.

Torgo, L., (2016). “Data mining with R: Learning with case studies.” New York: CRC Press.

Torgo, L., Branco, P., Ribeiro, R. P. and Pfahringer, B., (2015). “Resampling strategies for regression.” *Expert Systems*, 32(3), 465-476.

Veshosky, D., Beidleman, C. R., Buetow, G. W. and Demir, M., (1994). “Comparative analysis of bridge superstructure deterioration.” *Journal of Structural Engineering*, 120(7), 2123-2136.

Wellalage, N. K. W., Zhang, T., Dwight, R. and El-Akruti, K., (2013). “Bridge deterioration modeling by Markov Chain Monte Carlo (MCMC) simulation method.” In: P. W. T. Tse, J. Mathew, K. Wong, R. Lam, and C. N. Ko, Eds. *Proc. 8th World Congress on Engineering Asset Management and 3rd International Conference on Utility Management and Safety*, Hong Kong 30 October - 1 November 2013, 545-556.

Yamaguchi, T. and Hashimoto, S., (2010). “Fast crack detection method for large-size concrete surface images using percolation-based image processing.” *Machine Vision and Applications*, 21(5), 797-809.

Zhou, L., (2013). “Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods.” *Knowledge-Based Systems*, 41, 16-25.