An Ensemble Learning-Based Method for Missing Data Imputation in Road Maintenance

Shuyuan Tang

School of Economics and Management, Chongqing Jiaotong University, Chongqing, China
1494735337@qq.com

Abstract. To improve the accuracy of imputing missing maintenance engineering data, an ensemble learning method integrating three algorithms—XGBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—is proposed. This method constructs three base learners (XGBoost, SVM, and MLP) to establish the nonlinear mapping relationship between maintenance measures and multi-dimensional influencing factors through supervised learning. A soft voting ensemble strategy based on probability weighting is adopted to optimize the comprehensive decision-making effect of the model output, and the imputation performance of each model is systematically evaluated. The research results show that the proposed ensemble learning method achieves an accuracy of 99% in missing data imputation, which is significantly superior to the single models MLP (54%), SVM (72%), and XGBoost (77%). This verifies the effectiveness and superiority of the method in imputing missing maintenance data.

Keywords: highway pavement, pavement management data, data imputation, ensemble learning, road maintenance, base learner

1. Introduction

In recent years, with the increase in service life, asphalt pavements in China have entered a large-scale maintenance stage. Various degrees of damage have widely occurred, seriously affecting the pavement performance, and scientific and reasonable maintenance measures are urgently needed. When formulating maintenance decisions, it is necessary to make full use of historical pavement management data, and the quality of pavement management data directly affects the accuracy of decisions. Pavement management data mainly include five types of data—pavement structure materials, traffic volume, climatic environment, annual pavement condition detection, and maintenance engineering data. Among them, pavement condition detection data are mostly collected by multi-functional detection vehicles. Although certain deviations may occur due to vehicle or equipment offset during the detection process, the overall data quality is relatively reliable; maintenance engineering data mainly rely on manual daily records, which may have human errors such as omissions or miscalculations. The formulation of pavement maintenance measures is usually based on factors such as the previous year's pavement condition, traffic volume, pavement age, and climatic conditions. To fill the missing information in maintenance engineering data, it is necessary to establish a correlation model between it and data such as pavement condition, traffic volume, and

climatic environment. Therefore, before analysis, the pavement condition detection data should first be properly corrected to improve the reliability and consistency of the overall data.

In terms of data correction, Liu Tongbin [1] proposed that for a small amount of discontinuous missing values, fixed values, nearest neighbor imputation, and interpolation methods can be used for supplementation; continuous missing values are not processed; outliers can be handled by direct deletion, being regarded as missing values, and mean correction. Xiao [2] et al. used the K-Nearest Neighbour method for interpolating outliers and missing values in pavement management data. Guo et al. [3] eliminated unreasonable data using the interquartile range method. Han [4] proposed a data cleaning framework for asphalt pavement detection data based on neural networks, using artificial neural networks to clean abnormal data.

In terms of data imputation, Gao et al. [5,6] successively established maintenance engineering data monitoring models based on Bayesian models and Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) combined models (CNN-LSTM) to impute detected missing items; they also found that CNN-LSTM can capture the temporal characteristics of single pavement condition indicators and the spatial attributes of multiple pavement condition indicators, achieving a maximum accuracy of 87.5% without any feature extraction process. However, this accuracy is not sufficient as a reliable basis for maintenance decision-making. Zhang Haijiao [7] proposed a set of methods for identifying outliers and imputing missing values in maintenance data based on neural networks. He eliminated maintenance measures in abnormal data, and used normal data and neural network theory to construct the nonlinear mapping relationship between maintenance measures and other pavement factors to fill appropriate maintenance measures for abnormal data. The results showed that the proportion of abnormal data in pavement maintenance data decreased from 42.75% to 8.47%. Xiao Feng [8] proposed a method for imputing missing maintenance engineering data based on neural network models, and established a comparison between multi-class Logistic regression and maintenance engineering imputation models. The results showed that the prediction accuracy of the maintenance engineering imputation model based on neural networks was 98.02%, which was 17.63% higher than the latter. Huang Guodong [9] proposed a data cleaning method based on support vector machines for water supply pipe networks, which was verified with a case study of a certain city. The results showed that the data cleaning method based on support vector machines can effectively identify and repair outliers. Lu Xin [10] conducted data cleaning on the relevant characteristics of pavement smoothness based on the eXtreme Gradient Boosting (XGBoost) network, and predicted pavement smoothness combined with maintenance history. The results showed that this method has high prediction accuracy.

In summary, existing studies have proposed a variety of effective methods in data correction, including statistical imputation, K-nearest neighbor method, and neural network frameworks. However, there are still obvious deficiencies in the specific field of maintenance engineering data imputation: although complex models such as CNN-LSTM and logistic regression have been applied and shown potential, their accuracy is not as good as that of neural networks, and the effectiveness of some models that perform well in other fields in maintenance data imputation has not been verified, nor is there a systematic performance comparison with neural network models. More importantly, current research lacks an integration strategy that can effectively integrate the advantages of different models when the imputation effect of a single model is not good. To address the above problems, this paper proposes an ensemble learning-based maintenance engineering data imputation method, integrating the advantages of three algorithms: XGBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP); constructs three base learners (XGBoost, SVM, and MLP) to establish the nonlinear mapping relationship between maintenance measures and multi-

dimensional influencing factors through supervised learning; adopts a soft voting ensemble strategy based on probability weighting to optimize the comprehensive decision-making of the outputs of these three models; through a large number of field test data experiments and comparing the performance indicators of each base learner and the ensemble model, it is shown that the model has good application value in the imputation of pavement maintenance engineering data, aiming to provide effective support for the improvement of pavement management data quality and scientific utilization.

2. Principles and methods of the cleaning model

Pavement condition data and maintenance data in pavement management data are interdependent and cannot be cleaned simultaneously. Moreover, the omission of maintenance engineering data is the most serious, followed by errors in pavement condition detection values. This paper uses the interpolation method to correct the abnormally decreasing or increasing values of pavement condition data in adjacent years; supplements the missing maintenance engineering data according to pavement condition data and other relevant pavement management data.

2.1. Correction model for pavement condition data

Pavement condition data are derived from the Comprehensive Intelligent Condition Survey (CiCS) system, and this paper only uses the Pavement Condition Index (PCI) as a representative indicator for analysis. Among the pavement condition indicators of the research section, the PCI data are complete and have a large variation range, which is conducive to training during maintenance engineering imputation. Therefore, the pavement condition indicator selected in this paper is PCI. It is found that the variation law of some PCI data is inconsistent with reality. To improve the accuracy of model identification, it is necessary to correct the abnormal PCI data.

In this study, the interpolation method is used to correct abnormal PCI data to improve data quality and the accuracy of subsequent model identification. The confirmation of abnormal PCI data is to determine a reasonable variation interval through reliability analysis. The value with an improvement rate of 5% (95%) is selected as the lower (upper) limit of PCI improvement, that is, under the maintenance measure, there is a 95% (5%) probability that the PCI improvement value is greater (smaller) than a certain number. If a data point is outside this range, it is considered an abnormal data point. This study uses the linear interpolation method to calculate the corrected PCI value y(t) for the intermediate year t (ti < t < ti+1), and its calculation formula is:

$$y\left(t\right) = yi + \frac{yi+1-yi.}{ti+1-ti}\left(t - ti\right) \tag{1}$$

2.2. Imputation model for pavement maintenance data

The formulation of pavement maintenance measures is usually determined based on factors such as the previous year's pavement condition, traffic volume, pavement age, and climatic conditions. Therefore, before imputing missing maintenance engineering data, it is necessary to establish the mapping relationship between the type of maintenance engineering and variables such as the previous year's PCI, traffic volume, and climatic environment. This study uses three models—MLP, SVM, and XGBoost—to fit the mapping relationship between the type of maintenance engineering and relevant factors; the ensemble learning model adopts a soft voting ensemble strategy based on

probability weighting to optimize the comprehensive decision-making effect of the outputs of these three models.

2.2.1. Imputation model based on Multilayer Perceptron (MLP) algorithm

The imputation model based on the Multilayer Perceptron (MLP) algorithm has a fully connected structure, consisting of three parts: an input layer, a hidden layer, and an output layer.

According to the historical traffic volume and PCI data provided by enterprises and the climatic data from the China Meteorological Data Network, different explanatory variables have different ranges. Therefore, each explanatory variable needs to be standardized, as shown in Table 1. Among them, the total number of high-temperature days refers to the number of days in a year with an average daily temperature higher than 30°C; the total number of low-temperature days refers to the number of days in a year with an average daily temperature lower than 0°C; pavement age refers to the duration from the construction year to the data collection year.

Table 1. Data standardization methods

Feature Name	Original Range	Processing Method	Processing Formula
Traffic Volume (X1)	500-30000 vehicles/day	Dimension Normalization	X1/1000
Total Low-Temperature Days (X2)	0-100 days	Linear Scaling	X2/10
Total High-Temperature Days (X3)	0	Removal	—
Annual Precipitation (X4)	0-10000 mm	Removal	—
Previous Year's PCI (X5)	60-100	Centralization	X5-60
Current Year's PCI (X6)	60-100	Centralization	X6-60
Pavement Age (X7)	10-15 years	Dimension Compression	X7/10
PCI Difference (X8)	-20-20	Weight Enhancement	X8*2

The hyperparameters of the neural network include the number of hidden layers, the number of neurons in the hidden layers, activation function, loss function, batch size, number of epochs, learning rate, optimizer, and weight initialization method [7]. In this study, four different combinations of hidden layer structures are set: (32, 32) - two hidden layers, (64, 64) - two hidden layers, (64, 16, 64) - three hidden layers, and (32, 64, 32) - three hidden layers. The number of neurons in the hidden layer is approximately the sum of 2/3 of the number of neurons in the input layer and the number of neurons in the output layer [11]. The activation functions are ReLU, Tanh, and Logistic (Sigmoid).

Neuron output calculation: For a certain neuron in the hidden layer or output layer, its output is:

$$a_j = \sigma \left(\sum\nolimits_i w_{ij} . x_i + b_j \right)$$

(2)

Where: x_i is the input feature or the output of the neuron in the previous layer; w_{ij} is the connection weight; b_j is the bias term; σ is the activation function.

Loss function:

$$L = -\,_k y_k log \sum \left(y^k\right) \tag{3}$$

Where yk is the true label, and ŷk is the model prediction probability.
Optimization algorithm: SGD (Stochastic Gradient Descent):

$$w = w - \eta \cdot \nabla L \tag{4}$$

Adam is an optimization algorithm combining momentum and adaptive learning rate. It combines pavement maintenance measures, their corresponding influencing factor data, and neural network theory, determines the model results based on the grid search method, and constructs a pavement maintenance engineering data imputation method.

2.2.2. Imputation model based on Support Vector Machine (SVM) algorithm

For pavement maintenance engineering data Y (y1, y2, …, yt), support vector machine is used for regression fitting. Independent variables X (x1, x2, …, xl) are selected according to the strength of correlation. First, find a mapping φ(x) to map the input sample space to a high-dimensional feature space, and then use a linear function for regression. The regression function is shown in Equation (5):

$$f\left(x\right) = w^{\top} \cdot \varphi\left(x\right) + b \tag{5}$$

Where b is the threshold; w is the regression parameter.

According to the structural risk minimization principle of statistical learning theory, find the optimal w and b, and the optimization problem is shown in Equation (6):

$$\begin{cases} \min \ \frac{1}{2} w^{\top} w + C \sum_{i=1}^{l} \left(\xi_i + \zeta_i\right) \\ s.t. \ \ y_i - w \cdot \varphi\left(x_i\right) - b \leq \varepsilon + \xi_i \\ \ \ \ \ \ w \cdot \varphi\left(x_i\right) + b - y_i \leq \varepsilon + \zeta_i \end{cases}$$

$$\xi_i, \zeta_i \geq 0, \quad i = 1,2,\ldots,l \tag{6}$$

Where C is the penalty parameter; ε is the loss function, an inherent parameter of the support vector machine; ξ_i, ζ_i are slack variables that control the fitting error exceeding the precision. To solve the above equation, the Lagrangian function is introduced, as shown in Equation (7):

$$L\left(w, b, \xi_i, \zeta_i\right) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \left(\xi_i + \zeta_i\right) - \sum a_i \left(\varepsilon + \xi i + y i + w \cdot \varphi\left(x i\right) - b\right)$$

$$- \sum a_i^* \left(\varepsilon + \zeta i + yi + w \cdot \varphi \left(xi \right) + b \right) - \sum \left(\eta i \xi i + \eta i^* \zeta i \right) \tag{7}$$

Where αi, αi*, ηi, ηi* are Lagrange multipliers. To obtain the minimum value, the partial derivatives of the parameters w, b, ξ_i, ζ_i should all be zero. The convex optimization problem is transformed into Equation (8):

$$\min \ \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \left(a_i - a_i^* \right) \left(a_j - a_j^* \right) \cdot K \left(x_i, x_j \right) + \epsilon \sum_{i=1}^l \left(a_i + a_i^* \right) - \sum_{i=1}^l y_i \left(a_i - a_i^* \right) \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^l \left(a_i - a_i^* \right) = 0$$

$$0 \leq a_i, a_i^* \leq C$$

Where K($x_i \cdot x_j = \varphi \left(x_i \right) \varphi \left(x \right)$) is the kernel function.

Commonly used kernel functions include: radial basis function, polynomial function, linear function, etc. According to the correlation between independent variables and dependent variables of the monitoring data, this paper adopts the linear kernel function. The partial parameters where ($a_i - a_i^*$) is not zero are the support vectors (SV) in the problem. The linear regression estimation function obtained through learning is shown in Equation (9):

$$f \left(x \right) = \sum_{x_i \in SV} \left(\alpha_i - \alpha_i^* \right) \cdot K \left(x_i, x \right) + b \tag{9}$$

2.2.3. Imputation model based on XGBoost algorithm

XGBoost (eXtreme Gradient Boosting) is an efficient gradient boosting tree algorithm, and its core objective function consists of a loss function and a regularization term. For multi-classification problems, the model is constructed by optimizing the following objective function:

$$Obj \left(\theta \right) = \sum_{i=1}^n L \left(y_I, \widehat{y_I} \right) + \sum_{k=1}^k \Omega \left(fk \right) \tag{10}$$

Where: $L \left(y_i, \widehat{y_i} \right)$ is the multi-class cross-entropy loss function; $p_c \left(x_i \right)$ is the class probability output by the Softmax function; $\Omega \left(f_k \right)$ is the regularization term; T is the number of leaf nodes; w is the leaf weight; γ, λ, α are regularization coefficients.

Gradient boosting process:

The model constructs K trees through additive training iteration:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta f_t\left(x_i\right) \tag{11}$$

Where η is the learning rate, and each tree f_t is generated through a greedy algorithm. When splitting nodes, the gain is maximized:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{12}$$

Where $G = \sum g_i$ and $H = \sum h_i$ are the first and second derivatives of the loss function respectively; parameter optimization uses grid search (GridSearchCV) to find the optimal hyperparameter combination in the predefined space:

$$\Theta_{\text{opt}} = \underset{\theta\in\text{param\_grid}}{\arg\max} \text{Accuracy}_{\text{CV}} \tag{13}$$

The parameter space includes key parameters such as learning rate η, number of trees K, and tree depth d. Five-fold cross-validation is used to ensure the generalization ability of the model.

2.2.4. Construction of the ensemble learning model

This study adopts an ensemble learning strategy to improve classification accuracy. A single classification algorithm may lead to deviations in evaluation results due to different adaptability to different features of the dataset [12]. The ensemble learning model combines three different classifiers—XGBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—to further improve the classification accuracy of the model. The ensemble learning strategy gives the final result by integrating the training effects of different base learners. The commonly used voting methods for models include the hard voting strategy of majority voting and the soft voting decision based on the weighted average of classification probabilities. Ensemble prediction is performed through soft voting or hard voting. This method aims to utilize the advantages of different classifiers and improve the overall classification performance.

Soft voting integrates based on the probability distribution predicted by each classifier. A specific weight w_i is assigned to each classifier Ci, T is the number of classifiers, C_i^j represents the probability that the i-th classifier predicts to belong to class j, and $H^j\left(x\right)$ represents the probability that x belongs to class j. The final output is shown in Equation (14), and Equation (15) is the predicted label $\hat{y}$ of the click behavior.

$$H^j\left(x\right) = \frac{1}{T}\sum_{i=1}^{T} w_i C_i^j\left(x\right) \tag{14}$$

$$\hat{y} = arg \max \left[H^0 \left(x \right), H^1 \left(x \right) \right] \tag{15}$$

Hard voting is based on the majority vote of the predicted classes of each classifier. For sample X, the final predicted class $\hat{y}$ is determined by the following formula:

$$\hat{y} = mode \left\{ h_1 \left(x \right), h_2 \left(x \right), h_3 \left(x \right), \ldots, h_M \left(x \right) \right\} \tag{16}$$

Where $h_M \left(x \right)$ is the predicted class of sample X by the M-th classifier; $mode$ means taking the mode.

Comparing the two voting strategies, compared with soft voting, hard voting only considers classification labels and ignores the confidence of each model in the class, which may lead to information loss; when the probabilities predicted by the base learners are close, hard voting cannot effectively distinguish the uncertainty of the model, which may reduce the error robustness of the ensemble model. Therefore, soft voting can not only improve the generalization ability of the ensemble model but also make the prediction result closer to the statistical optimal solution.

3. Imputation process of pavement maintenance data

3.1. Data imputation process

First, the entire original data are randomly divided into a training set and a test set in an 8:2 ratio; among them, the MLP model uses 10% of the training data as a validation set to prevent overfitting; the three base classifiers are trained separately, and each model undergoes an independent hyperparameter optimization and cross-validation process. The ensemble learning model loads and processes the training data and prediction data, and determines the final class through soft voting.

The MLP model first initializes an MLP classifier with an early stopping mechanism, with a maximum number of iterations of 2000. By reserving 10% of the training data as a validation set, dynamic monitoring of the training process is achieved to prevent overfitting; the performance of different hyperparameter combinations is systematically evaluated through grid search combined with five-fold cross-validation (cv=5); the accuracy rate is used as the optimization index to ensure the statistical reliability of parameter selection. After determining the optimal parameters, the model is retrained with the best configuration, and finally, prediction is performed on the independent test set. Eventually, the optimal parameters of the model are: the hidden layer adopts (64, 64), the activation function adopts 'relu', and the optimization algorithm adopts 'sgd'.

The SVM model first initializes an SVC classifier, sets probability=True to enable the probability estimation function; sets a parameter grid including regularization parameter C (10/50/100), kernel function ('linear', 'poly', 'rbf'), kernel coefficient ('scale', 'auto'), and polynomial order (2, 3); through five-fold cross-validation, searches for the optimal parameter combination with accuracy as the evaluation index. After obtaining the optimal parameters and the corresponding validation score, the final model is retrained with the optimal configuration. Eventually, the optimal parameters of the model are: the regularization parameter C is 10, the kernel function adopts 'poly', the kernel coefficient adopts 'auto', and the polynomial order is 2.

The XGBoost model follows the machine learning paradigm of Gradient Boosting Decision Tree (GBDT). First, initialize an XGB classifier, clarify the type and number of parameters, and define multi-class logarithmic loss as the evaluation index; construct a search space for 8 key hyperparameters: learning rate (0.01, 0.1, 0.001), number of trees (100, 200, 300), maximum depth (3, 5, 7), subsample rate (0.7, 0.8, 0.9), feature sample rate (0.7, 0.8, 0.9), L1/L2 regularization terms (0, 0.1, 0.5), and minimum leaf sample weight (1, 3, 5). Through GridSearchCV for five-fold cross-validation, after evaluating 6561 parameter combinations, the accuracy rate is used as the optimization target, and parallel computing is used to improve the search efficiency. After obtaining the optimal parameter combination, the best model is reconstructed and its performance is evaluated on the independent test set. Eventually, the optimal parameters of the model are as follows: learning rate is 0.1, number of trees is 100, maximum depth is 5, subsample rate is 0.9, feature sample rate is 0.7, L1 regularization parameter is 0, L2 regularization parameter is 0.1, and minimum leaf sample weight is 1.

3.2. Evaluation of imputation effect

The imputation effect of missing maintenance records is evaluated using four indicators: Accuracy, Precision, Recall, and F1-score. The following table shows their calculation formulas. True Positive (TP): the number of positive classes predicted as positive classes; True Negative (TN): the number of negative classes predicted as negative classes; False Positive (FP): the number of negative classes predicted as positive classes (Type I error); False Negative (FN): the number of positive classes predicted as negative classes (Type II error) [13].

Table 2. Calculation formulas of model evaluation indicators

Evaluation Indicator	Formula
Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad \frac{TP+TN}{TP+FN+FP+TN}$
Precision	$\text{Precision} = \frac{TP}{TP+FP}$
Recall	$\text{Recall} = \frac{TP}{TP+FN}$
F1-score	$F_1 = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recsll}}$

4. Case study of pavement management data cleaning

To verify the applicability of the proposed ensemble model in practical engineering applications, the pavement management data of a certain expressway in Shanxi Province are used for effect verification. This expressway is a two-way four-lane road. Pavement condition data are collected with one measuring point every 1 kilometer, and a total of 3680 Pavement Condition Index (PCI) detection values have been accumulated from 2015 to 2024. The time arrangement of pavement detection and maintenance is as follows: pavement condition detection is carried out every December, maintenance is implemented around June of the following year, and detection is carried out again in December of the same year, forming an annual cycle of "detection—maintenance—re-detection". Taking a section of this expressway as an example, the PCI records and corresponding maintenance measures of two pavement units from 2015 to 2024 are counted.

According to the pavement management data, there are multiple logical contradictions between the time series of pavement condition detection and the time series of maintenance measures. It is

speculated that data anomalies may come from the following three aspects: (1) missing manual maintenance records; (2) errors in existing maintenance records; (3) errors in PCI detection values.

4.1. Correction results of pavement condition data

To improve data quality, this study evaluates the rationality of PCI changes based on the concept of reliability. There are four types of historical maintenance measures for this section: surface regeneration repair, milling and resurfacing, ultra-thin wearing course overlay, and no maintenance. The distribution of PCI improvement values in the second year after the implementation of various measures at different percentiles is counted. Based on statistics, the 5th percentile value is taken as the lower limit of PCI improvement, and the 95th percentile value is taken as the upper limit to construct a reasonable interval for PCI changes under various maintenance measures, as shown in Table 3.

Table 3. Reasonable interval of PCI changes

Maintenance Measure	Lower Limit	Upper Limit
No Maintenance	-3.96	-31.48
Milling and Resurfacing	0.23	28.34
Ultra-thin Wearing Course Overlay	1.26	31.36
Surface Regeneration Repair	-3.45	8.84

For abnormal PCI samples outside the reasonable interval, this study uses the interpolation method for correction, and finally obtains the cleaned pavement condition data, providing a reliable basis for subsequent model verification.

4.2. Imputation results of maintenance engineering data

To effectively repair the problem of missing maintenance records in pavement management data, this study constructs an ensemble learning model that integrates three classifiers with different principles: XGBoost based on tree method, SVM based on kernel method, and Multilayer Perceptron (MLP). Through the ensemble strategy, the bias and variance of a single model can be effectively reduced, and the soft voting mechanism is used to integrate the prediction probabilities of each model to improve the overall classification performance.

The results show that the proposed ensemble learning model achieves an accuracy of 99% in the maintenance measure classification task, which is significantly superior to each single model. Among them, the accuracy rates of MLP, XGBoost, and SVM are 54%, 72%, and 77% respectively. As shown in Figure 1, a further comparison of the performance of each model on multiple evaluation indicators shows that the ensemble model is significantly superior to any single model in all indicators.
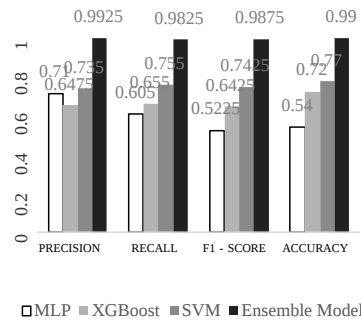
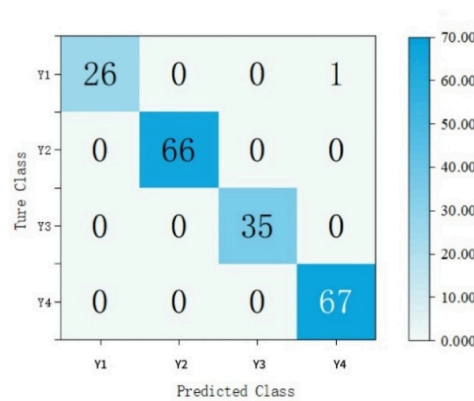Figure 1. Comparison of imputation effect evaluation indicators of each model

Figure 2. Confusion matrix of maintenance measure prediction results for test samples

This study uses the ensemble model to impute missing maintenance engineering data. Figure 1 shows the detailed evaluation results on four types of maintenance (no maintenance, surface regeneration repair, milling and resurfacing, ultra-thin wearing course overlay). It can be seen from the figure that the accuracy, precision, recall, and F1-score obtained on the four types of samples. The results show that the F1-score of all classes is higher than 0.95, indicating that the ensemble learning model has good and stable data repair ability in maintenance engineering data imputation. For 195 test samples, the number of pavement units whose predicted maintenance measure type is the same as the actual maintenance measure type is 194 (as shown in Figure 2).

5. Conclusion

Aiming at the frequent occurrence of missing maintenance engineering records in pavement management data, the ensemble learning algorithm proposed in this paper systematically evaluates the data imputation effect of the model from multiple dimensions such as accuracy, precision, and recall. The main conclusions are as follows:

(1) In the process of pavement condition data cleaning, the reasonable intervals of PCI changes corresponding to four maintenance measures are determined through the statistical percentile method. The lower limits of PCI changes for no maintenance, surface regeneration repair, milling and resurfacing, and ultra-thin wearing course overlay are -3.96, -3.45, 0.23, and 1.26 respectively, and the upper limits are -31.48, 8.84, 28.34, and 31.36 respectively.

(2) Single models have limited performance in maintenance data imputation tasks: the average accuracy rates of MLP, SVM, and XGBoost are 54%, 72%, and 77% respectively; the average F1-scores are 0.54, 0.74, and 0.64 respectively; the average recall rates are 0.605, 0.755, and 0.655 respectively; the average precision rates are 0.701, 0.735, and 0.648 respectively.

(3) The ensemble learning model proposed in this paper is significantly superior to single models in all evaluation indicators. Its comprehensive accuracy rate reaches 99%, and the average values of F1-score, recall rate, and precision rate are 0.9875, 0.9825, and 0.9925 respectively, showing excellent maintenance data repair ability and engineering applicability.

The imputation method proposed in this study is a probability-based integration, which reduces the bias of a single model and has better tolerance for data noise and outliers. Moreover, the method in this study can not only be applied to the imputation of missing data but also to data prediction.

References

[1]   Liu, T. B., Yao, S. W., Xu, Z. X., et al. (2021). Research on network-level pavement maintenance decision scheme based on multi-data fusion analysis. Modern Transportation Technology, 18(3), 23-27+32.

[2]   Xiao, F., Chen, X., Cheng, J., et al. (2023). Establishment of probabilistic prediction models for pavement deterioration based on Bayesian neural network. International Journal of Pavement Engineering, 24(2), 2076854.

[3]   Guo, F., Zhao, X., Gregory, J., et al. (2021). A weighted multi-output neural network model for the prediction of rigid pavement deterioration. International Journal of Pavement Engineering, 23(8), 2631-2643.

[4]   Han, C., Zhang, W., & Ma, T. (2022). Data cleaning framework for highway asphalt pavement inspection data based on artificial neural networks. International Journal of Pavement Engineering, 23(14), 5198-5210.

[5]   Gao, L., Qiu, S., & Prasad, T. R. (2017). Bayesian detection of unrecorded maintenance and rehabilitation treatments in pavement management. Paper presented at the Transportation Research Board 96th Annual Meeting, Washington DC, United States.

[6]   Gao, L., Yu, Y., Hao, R. Y., et al. (2021). Detection of pavement maintenance treatments using deep-learning network. Transportation Research Record: Journal of the Transportation Research Board, 2675(9), 1434-1443.

[7]   Zhang, H. J. (2023). Research on outlier identification and missing value imputation methods for highway pavement maintenance data based on neural network. Highway, 68(5), 365-370.

[8]   Xiao, F. (2023). Research on maintenance management and optimization decision of expressway asphalt pavement (Unpublished doctoral dissertation). Southeast University.

[9]   Huang, G. D., Long, Z. H., Zhu, Z. P., et al. (2022). Water supply network monitoring data cleaning based on support vector machine. Water & Wastewater Engineering, 58(9), 124-129.

[10] Lu, X., & Qian, X. D. (2023). Research on pavement roughness prediction method based on XGBoost. Municipal Engineering Technology, 41(6), 10-14.

[11] Heaton, J. (2008). Introduction to neural networks with Java. Heaton Research, Inc.

[12] Zhao, Y. M., Liu, S. S., & Lü, L. C. (2025). Research on early identification of disruptive technologies based on ensemble learning—A case study in the field of quantum computing. Data Analysis and Knowledge Discovery, 1-21. https: //link.cnki.net/urlid/10.1478.G2.20250227.1446.006

[13] Zhang, T. J., & Han, H. H. (2021). Research on asphalt pavement crack identification and classification based on residual neural network. Highway, 66(10), 24-29.