J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 7321--7336

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Intelligent Estimation of Construction Project Costs Based on Subtractive Clustering-based Self-Learning Convolutional Neural Network

Ning Feng[1], ✉

[1] School of Management, Henan University of Urban Construction, Pingdingshan, Henan, 467036, China

ABSTRACT
Traditional construction project cost estimation methods rely on expert experience and statistical models, which are difficult to handle complex data and multimodal features effectively and have low prediction precision. This paper constructs an intelligent building engineering cost estimation model that combines subtractive clustering, a self-learning mechanism, and convolutional neural networks (CNN) to address this problem. In the data preprocessing stage, subtractive clustering is applied to optimize multimodal data, screen key features, and eliminate redundant information. Subsequently, the model parameters are dynamically adjusted according to the error feedback through a self-learning mechanism to improve its adaptability to diverse construction projects. In the feature extraction and estimation stage, the CNN module is combined to extract deep features from images, texts, and numerical data to achieve high-precision estimation. The experimental results show that the model in this paper outperforms traditional methods in terms of MSE (mean-square error), MAE (mean absolute error), R^2 (coefficient of determination), MAPE (mean absolute percentage error), with the mean values being 73.18, 8.33, 0.9477, and 5.33%, respectively. In summary, the model in this paper demonstrates superior precision, adaptability, and robustness in construction project cost estimation.

Keywords: Subtractive Clustering Algorithm, Self-Learning Mechanism, Convolutional Neural Network, Construction Project, Cost Estimation

1. Introduction

There are many cases in the field of engineering construction in which there are obvious deviations between the actual investment amount and the estimated value due to insufficient understanding, and such deviations will not only cause great losses to investors and construction companies, but also affect the success and profitability of the whole project [1-3]. Construction project cost

✉ Corresponding author.
E-mail address: 30050701@huuc.edu.cn (N. Feng).
Received 05 March 2024; Revised 10 May 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-406

estimation is crucial for improving the core competitiveness of construction units. Through the construction cost estimation, the construction unit can accurately predict the investment scale of the construction project and determine the feasibility of the project [4]. Based on this, the construction unit can make a detailed project plan and time schedule to ensure the smooth implementation of the project [5-6]. In addition, through the construction cost estimation, the construction unit can predict and effectively control the cost of the project and improve the economic efficiency of the project [7-8]. However, cost estimation itself is full of challenges because the characteristics and conditions of engineering projects are diverse and full of uncertainties, which leads to the fact that accurately estimating construction cost has been a complex and difficult task [9-11]. Considering that the accuracy of the estimation results has a direct and far-reaching impact on the investment and decision-making of the project, the study of engineering cost estimation modeling is particularly important.

China's engineering cost system still plays a great role, while the new cost system is still not established, the study of cost control means of engineering construction enterprises will help to form a new engineering cost management mode in China, reduce the cost of engineering construction, save construction funds, and better utilize the social benefits of engineering construction [12-15]. Traditional cost estimation methods usually rely on statistical analysis and simple regression theory, however, these methods have the defects of low precision and long time-consuming, and the inability to measure the cost changes due to the adjustment of local programs or engineering features limits their application in complex engineering projects [16-20]. In recent years, with the rapid development of computer technology, new estimation methods have been introduced into cost estimation, providing new ways to improve accuracy and efficiency [21-23].

Nowadays, with the development of artificial intelligence technology, researchers have begun to apply techniques such as artificial neural networks, fuzzy mathematics and gray systems to the field of engineering cost estimation. Modern neural network is a kind of nonlinear statistical model usually used to establish complex relationships between inputs and outputs, or used to explore the patterns of data, and is increasingly used in engineering cost measurement.Jiang, Q. evaluated the performance of BP neural network in the application of cost estimation of construction projects, and its good estimation accuracy provides data support for the investment decision-making of the construction unit [24]. Du , Z. et al. proposed a project cost estimation model based on improved BP neural network, using genetic algorithm to optimize the model correlation coefficients, simplifying the process of calculating investment and cost of the project, and realizing fast and accurate project estimation [25]. Xu, X. et al. used the sparrow search algorithm to improve the BP neural network model for the prediction of the cost of grid project, which improves the uncertainty and complexity data The accuracy of estimation of engineering falsification under uncertainty and complexity data is improved to provide support for practical application research [26]. Xin, Z. combines BIM technology with BP neural network to construct a prediction model for engineering cost management, and the engineering quantity calculated by BIM technology is used as the input data for training, so as to realize accurate engineering cost estimation of the construction project at a low cost [27].

The use of fuzzy mathematical theory in project estimation can be carried out quickly, eliminating the need for cumbersome quantity calculations for substitute projects. Wang, X. showed that a construction project cost estimation model based on fuzzy mathematical methods can estimate the cost of the project to be estimated based on the data of the constructed project and shows good accuracy and applicability [28]. Habibi, F. et al. used the critical path method and the project Evaluation and Review Technique to determine the fuzzy numbers for project cost estimation and used a step-by-step algorithm to de-fuzzify in order to estimate the actual project cost [29]. Beilin, I. L. devised a model for estimating the cost of innovative projects based on discrete and continuous fuzzy numbers, which overcame the problem of the inability to mathematically describe the linguistic variables, and provided the decision makers with an accurate cost estimate through the use of cash flow analysis [30]. Obianyo, J. I. et al. utilized fuzzy logic computational tool to assess the factors associated with cost overruns in construction projects, and their cost estimation results showed high similarity between the cost estimation results and the actual costs with good predictive performance [31].

There are also scholars through the construction of gray measurement model can be estimated on the construction cost, in which the known historical data of construction cost, such as white parameters, there are some unknown data, such as the project management situation, market changes in the process of construction of new projects, which belong to the unknown gray parameters. Xu, Y. et al. introduced TSNE technology on the basis of gray correlation algorithm and used it to construct a construction project cost prediction model, which still has good prediction effect in the face of increasingly complex construction project cost prediction [32]. Nadafi, S. et al. proposed an Earned Value Management approach based on Gray Theory, which allows construction project managers to make accurate predictions of budgets and time under conditions of uncertainty [33]. Xie, S. et al. established an engineering cost prediction model integrating gray theory and BP neural network for situations such as missing sample data or insufficient data volume, and input construction materials and mechanical displacement data into the neural network for training to realize accurate prediction of bad data [34]. Jia, H. examined the effectiveness of a gray Markov chain optimization model applied to project cost estimation, which can accurately identify construction project budget risks and simplify the cost estimation process, which is highly useful in construction project management [35].

Although the above engineering estimation model has achieved satisfactory results, its accuracy rate still cannot meet the actual project management needs. With the proposal of deep learning theory and the improvement of computing performance of hardware devices, convolutional neural network has gained great success in the fields of computer vision and natural language processing with higher recognition rate, which is introduced into the application of construction project cost estimation in order to further improve the accuracy of cost estimation. Although some existing studies have promoted the development of construction project cost estimation methods to a certain extent, there are still problems, such as strong dependence on feature extraction, insufficient dynamic adaptability, and limited processing effect on complex multimodal data. In this regard, this paper attempts to combine subtractive clustering, self-learning mechanism, and CNN to overcome these limitations and further improve the ability of multimodal data fusion and precise estimation, providing new ideas and technical support for construction project cost estimation.

2. Construction Project Cost Estimation Model Construction

2.1. *Overall Model Architecture Design*

The construction project cost estimation model constructed in this paper consists of three main modules: subtractive clustering module, self-learning mechanism module, and CNN module. The modules cooperate to form an organic overall architecture. Figure 1 presents the specific structure:

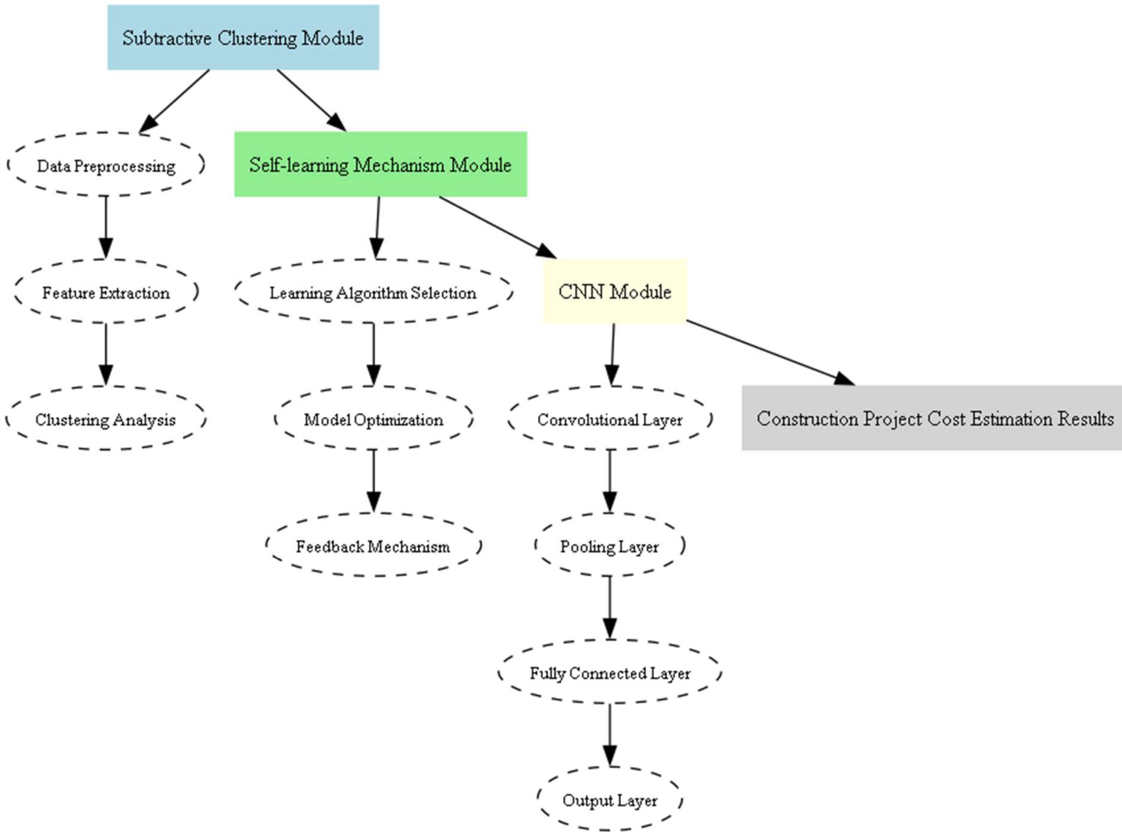

Fig. 1. Construction project cost estimation model structure

1) Subtractive clustering module. This module is mainly used for data preprocessing and feature optimization. The model can extract highly representative and practical information from complex multimodal data through the subtractive clustering algorithm. Subtractive clustering can effectively identify potential patterns in data through density calculation, eliminate redundant features, and provide concise and useful input for subsequent feature extraction and model training.

2) Self-learning mechanism module. The core function of the self-learning mechanism module is to achieve dynamic optimization of the model. This module gradually improves the generalization ability and adaptability of the model through continuous error feedback and parameter adjustment. It enables the model to automatically adjust the learning strategy and model parameters according to new input data in different construction projects to ensure that a high prediction accuracy is maintained in a changing project environment.

3) CNN module. The CNN module is this model's core feature extraction module, responsible for extracting deep features from multimodal data. CNN efficiently processes images, texts, and numerical data in construction projects through convolutional and pooling layers. In image data processing, CNN can identify key details in architectural design drawings and construction drawings. In text and numerical data processing, project potential patterns are captured through multi-layer convolutional structures. This module provides a solid feature foundation for the final estimation.

2.2. Data Preparation and Subtractive Clustering Optimization

1) Data collection. The data for this study mainly comes from multiple actual construction projects involving construction projects of different types and scales. The collected data covers information in various dimensions, including the project's design drawings, construction drawings, budget reports,

historical cost data, construction progress, and material costs. The data is from cooperative construction companies, public bidding documents, and related industry databases, ensuring the diversity and representativeness of the data. Specifically, the collected data can be divided into image data, text data, and numerical data. Image data includes architectural design drawings, construction drawings, and renderings. The amount of image data for each project is between 1,000 and 2,000. Text data includes project contract documents, budget instructions, construction progress reports, etc., and each project contains 30 to 50 documents. Numerical data covers material unit prices, labor costs, construction periods, etc., with about 300 to 500 records. By collecting multi-dimensional data from different types of construction projects, data from about 50 projects is collected, totaling about 100,000 data points, ensuring the representativeness and comprehensiveness of the data.

2) Data preprocessing. In the data preprocessing process, this study first performs data cleaning to ensure the quality and consistency of the input data. For all data sources, missing values and outliers are first removed, especially the missing or erroneous records in the numerical data. For the detection of outliers, the interquartile range (IQR) method is used for processing, and its calculation formula is:

$$IQR = Q_3 - Q_1 \tag{1}$$

Among them, Q_1 and Q_3 are the first and third quartiles, respectively. The criterion for determining outliers is:

$$X < Q_1 - 1.5 \cdot IQR \; or \; X > Q_3 + 1.5 \cdot IQR \tag{2}$$

For some missing data, the integrity and accuracy are maintained by interpolation or deletion of outliers. Regarding image data, the input images are standardized, including cropping, rotation, and scaling, to ensure that the image size and resolution meet the input requirements of the model. The pixel values of the image are normalized to the range of [0,1], and the normalization formula is:

$$I_{scaled} = \frac{(I - I_{\min})}{I_{\max} - I_{\min}} \tag{3}$$

Among them, I is the original pixel value, and $I_{\min}$ and $I_{\max}$ are the image's minimum and maximum pixel values, respectively. In addition, automated tools are used for design and construction drawings to remove blank areas at the edges and uniformly scale the images to a fixed size to ensure that the image details are not distorted, providing high-quality input for subsequent feature extraction and model training.

Natural language processing (NLP) technology is used for text data to preprocess project contracts, budget instructions, and other documents. First, the text is divided into basic phrase units through word segmentation technology, and meaningless information such as stop words and punctuation marks are removed. Then, the TF-IDF (Term Frequency-Inverse Document Frequency) method is used to extract keywords, and the calculation formula is:

$$TF - IDF(t,d) = TF(t,d) \cdot IDF(t) \tag{4}$$

Through this method, the core information related to the construction project cost is screened out. For numerical data, the standardization method is used for processing. The standardization formula is:

$$X_{\text{standard}} = \frac{X - \mu}{\alpha} \tag{5}$$

Among them, μ is the mean, and α is the standard deviation. This method can scale the data of indicators such as construction period and material cost to a uniform magnitude, avoiding the deviation of model training caused by the enormous value range of some features.

2) Application of subtractive clustering algorithm. In the data preprocessing stage, the subtractive clustering algorithm is vital in optimizing features and eliminating redundant information. Figure 2 shows its working principle:

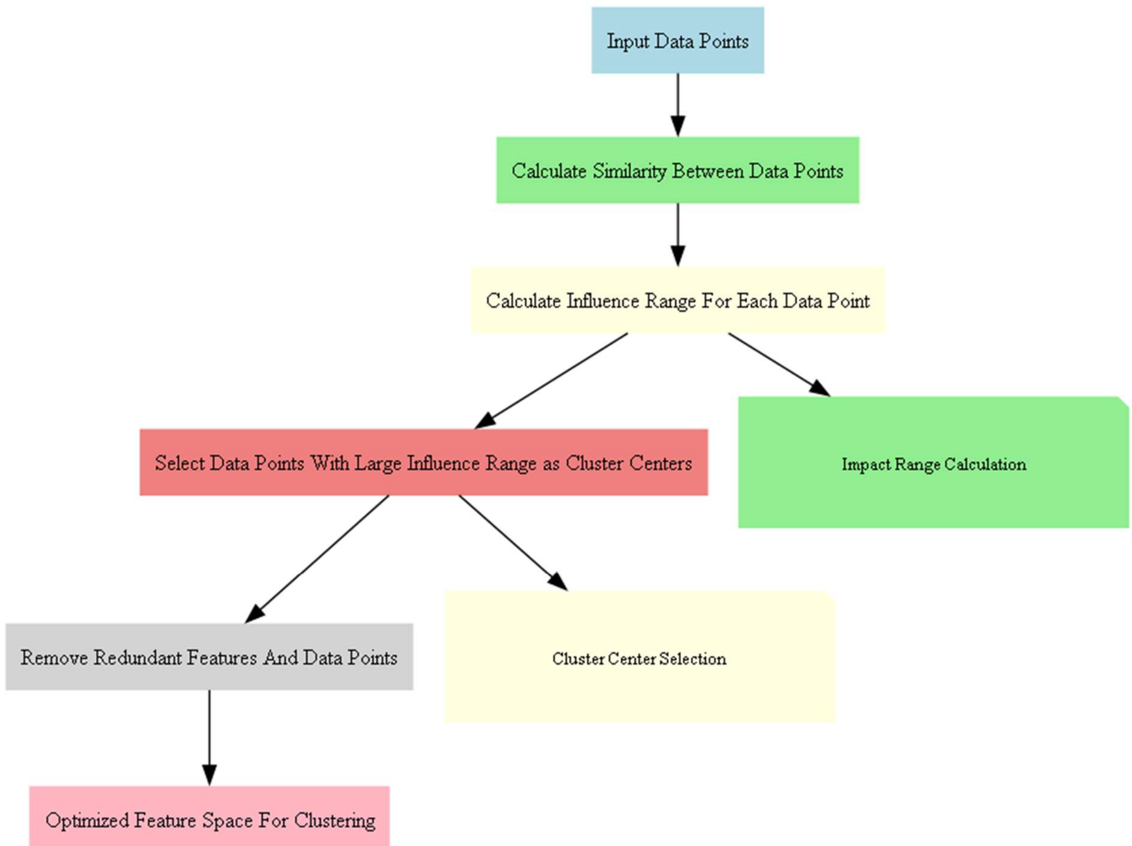

Fig. 2. Subtractive clustering algorithm work principle

The algorithm calculates the similarity between data points, evaluates the “influence” of each data point in the data set, and selects representative data points as cluster centers based on this influence. Specifically, the algorithm calculates an “influence range” for each data point, and points with an extensive influence range are considered essential representatives in the data set. The calculation formula for the influence range is:

$$r_i = \sum_{j=1}^{N} exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{6}$$

Among them, r_i is the influence range of data point x_i. $\left\|x_i - x_j\right\|$ is the Euclidean distance between data points x_i and x_j. σ is the width parameter of the Gaussian kernel. N is the total number of data points in the data set. The essence of the subtractive clustering algorithm is to automatically remove redundant features through density calculation and retain the most representative information, thereby reducing noise in the data and optimizing the subsequent feature extraction and modeling process.

In this study, the subtractive clustering algorithm is applied to the processing of multimodal data, especially in optimizing feature space. First, the image data, text data, and numerical data from different data sources are preliminarily integrated and feature-extracted to form a relatively broad feature space. Subsequently, the subtractive clustering algorithm optimizes the feature space by calculating the influence of each data point. The algorithm selects data points with an extensive influence range as cluster centers in this process. The selection formula of cluster centers is:

$$c = \arg\max ri \tag{7}$$

Points with a negligible impact range are removed, thereby reducing redundant features. In image data of construction projects, subtractive clustering algorithms can identify key design details in images and remove unnecessary background or noise areas. The algorithm selects the most representative core information from contract documents and budget instructions for text data by evaluating the density of vocabulary and sentences. For numerical data, subtractive clustering can optimize the feature space and remove redundant or irrelevant information by selecting cost data or construction period data with a significant impact range as cluster centers.

2.3. Self-Learning Mechanism and Dynamic Optimization

1) Mechanism principles. The self-learning mechanism adjusts model parameters through dynamic feedback, aiming to improve the model's adaptability to different types of construction projects. Figure 3 shows the implementation process of this mechanism:

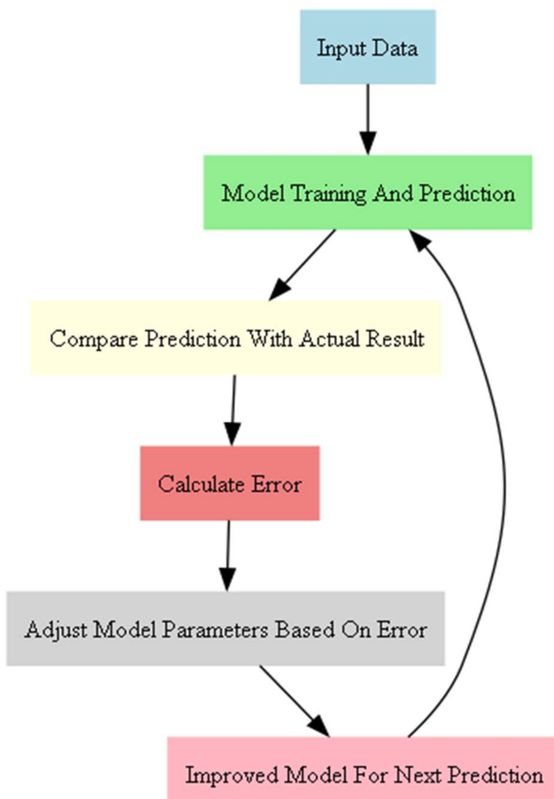

Fig. 3. Self-learning mechanism implementation process

The core idea of this mechanism is to continuously adjust the model parameters according to the error feedback information during the model training and prediction process to improve the prediction precision and robustness of the model. In the process of construction project estimation, the data is highly diverse, and the characteristics and environment of the project change dynamically over time. For example, factors such as the project's scale, budget, and construction methods may change over time and under different environmental conditions, making it difficult for traditional static models to cope with these changing data. Comparatively, self-learning mechanisms enable real-time adjustment of the model so that it always remains adaptable to the latest data.

In the self-learning mechanism, the model compares each prediction with the actual result, calculates the error, and adjusts the parameters to reduce the deviation in future predictions. This continuous adjustment process enables the model to maintain efficient prediction capabilities in different construction projects.

2) Optimization process. The self-learning mechanism improves the model's generalization ability and robustness by dynamically optimizing the model through the steps of error feedback, weight adjustment, and learning rate regulation. Error feedback is its core. The error is calculated and fed back to the model for adjustment by comparing the predicted results with the actual values. The error is quantified by mean-square error (MSE):

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{8}$$

Among them, y_i is the actual value, and $\hat{y}_i$ is the model's predicted value. The model adjusts the weights using the backpropagation algorithm based on the error feedback. The weight update is expressed by Formula (9):

$$\theta_i{}' = \theta_i - \eta \cdot \frac{\partial L}{\partial \theta_i} \tag{9}$$

Among them, θ_i is the i-th parameter of the model. $\theta_i{}'$ is the adjusted parameter. η is the learning rate. $\frac{\partial L}{\partial \theta_i}$ is the partial derivative of the loss function concerning parameter θ_i.

The learning rate is a key hyperparameter in model updating, which determines the magnitude of each parameter adjustment. The learning rate needs to be dynamically adjusted in the self-learning mechanism according to the error changes. If it is too large, it may lead to skipping the optimal solution, while if it is too small, it may reduce the training efficiency. Therefore, the learning rate adjustment mechanism automatically adjusts according to the error changes to ensure the stability and efficiency of the model. Common methods include learning rate decay and adaptive adjustment. Learning rate decay follows Formula (10):

$$\eta(t) = \frac{\eta_o}{1 + \lambda t} \tag{10}$$

$\eta(t)$ is the learning rate at the t-th iteration. η_o is the initial learning rate. λ is the decay factor. As the training progresses, the learning rate gradually decreases, allowing the model to converge stably when approaching the optimal solution.

2.4. Feature Extraction and Estimation of Convolutional Neural Networks

1) CNN module design. CNN is a model widely used in deep learning to process images and other structured data. It automatically extracts features from data by simulating the human visual system. For the construction cost estimation task, this study designs a multi-level CNN module to process multimodal data such as images, texts, and numerical values in construction projects, extracting deep features and building an effective estimation model.

The core of the CNN module consists of a stack of multiple convolutional layers and pooling layers. This study selects a 3×3 convolution kernel for feature extraction in the convolutional layer. The size of the convolution kernel is 3×3, which can effectively capture local features in a smaller receptive field and is suitable for structural information in image data such as architectural design drawings and construction drawings. The stride of the convolution operation is set to 1, which means that each convolution operation covers every pixel of the input image, thereby ensuring complete extraction of detailed information.

Following the convolutional layer is the pooling layer, mainly used for dimensionality reduction and feature compression. The pooling operation reduces the data size by selecting the maximum or average value in the convolutional feature map while maintaining the most representative features. This study selects the maximum pooling method to reduce the dimension and retain critical information effectively. The size of the pooling kernel is 2×2, and the step size is set to 2. This setting not only helps to compress the size of the feature map but also reduces the computational complexity and effectively prevents the overfitting of the model.

2) Multimodal feature fusion. Construction project cost estimation requires comprehensive consideration of information from multiple data modalities. Therefore, multimodal feature fusion is crucial to improving the model's estimation ability. This study uses a fully connected layer to effectively fuse features of different modalities, such as images, text, and numerical values, to form a unified feature representation.

First, for image data, the deep feature map extracted by the CNN module is mapped to a high-dimensional feature space through a fully connected layer and further converted into a feature vector suitable for the estimation task. Text data is transformed into a numerical vector through NLP technology to represent the key information in the text information. Numerical data such as material cost and construction period are converted into a unified numerical format through standardized methods so that they can be fused with image and text features.

The fully connected layer maps data from different modalities into a unified feature space through a linear transformation so that information from different modalities can be processed in the same representation space. Specifically, assuming that the feature vectors of each modality are F_{image}, F_{text}, and $F_{numerical}$, the fused feature representation F_{fusion} can be expressed as a weighted sum:

$$F_{fusion} = w_1 \cdot F_{image} + w_2 \cdot F_{text} + w_3 \cdot F_{numerical} \tag{11}$$

Among them, w_1, w_2, and w_3 are the weights of each mode, reflecting the relative importance of different data sources in the model. Weighted fusion can not only maintain the information characteristics of each mode but also give full play to the role of each modality in construction project estimation.

3) Estimation output. After feature extraction and fusion, the estimation model in this study enters the end-to-end estimation process. This process generates the construction project cost estimation result through the output layer. Specifically, the fused feature representation F_{fusion} is further mapped through several fully connected layers. Finally, the cost estimation value of the construction project is output.

However, the estimation results are not entirely correct. Due to construction projects' complexity, the estimation model's results inevitably have certain errors. The sources of errors mainly include data noise, feature extraction limitations, and the model's generalization ability. This study uses indicators such as MSE and MAE to evaluate the model's output to assess its estimation precision. The calculation formula of MAE is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \tag{12}$$

2.5. Model Training and Deployment Application

1) Model training and validation. In model training and validation, this study first divides the data set according to a certain ratio to ensure the reasonable distribution of training data and validation data. Specifically, the data set is divided into an 80% training set and a 20% validation set. The training set is used to train the model. The validation set is used to evaluate the performance of the model to ensure that the model can maintain good performance on both seen and unseen data. The data division follows the principle of random allocation to avoid the bias in data division affecting the training effect of the model. In addition, to further enhance the robustness of the model, K-fold cross-validation (K=5) is applied to evaluate the stability and robustness of the model with different data divisions.

A mini-batch SGD optimizer is used during the training process. The batch size is 32 to balance computational efficiency and model convergence speed. The model training epochs are set to 100, and the early stopping mechanism is applied. When the performance on the validation set does not improve for 10 consecutive epochs, the training is stopped to prevent overfitting. The initial value of the learning rate is set to 0.0001, and the exponential decay strategy is adopted. The learning rate is decayed to 90% of the original value every 10 epochs to ensure that the model can more stably approach the optimal solution later in training.

In addition, to optimize the model's performance, this study uses grid search to adjust hyperparameters such as the number of filters, convolution kernel size, pooling kernel size, step size, number of nodes in the fully connected layer, and learning rate during training. Multiple training and validation are performed for different hyperparameter combinations. The optimal hyperparameter configuration is finally obtained as 64 convolution filters, 3×3 convolution kernels, 2×2 pooling kernels, a step size of 1, 512 fully connected nodes, and a learning rate of 0.001. With this configuration, the model's prediction accuracy and generalization ability are significantly enhanced.

2) Application deployment process. After model training, it enters the deployment and application stage of the actual construction project cost estimation scenario. The system designs an application programming interface (API) interface based on the RESTful architecture, which supports JSON (JavaScript Object Notation) format data transmission to ensure efficiency and scalability. The interface receives various input data, such as project drawings, cost estimation data, and construction plans. It ensures the validity of data format and content through an input verification mechanism while supporting asynchronous processing to meet high concurrency requirements.

The system develops a dynamic visualization function to improve the readability and practicality of the estimation results. The estimation results are displayed in dynamic charts and reports through the front end, including key indicators such as construction material cost distribution, labor cost, and construction period. The front-end interface built with D3.js and Plotly tools supports PDF (Portable Document Format) export and interactive dashboard functions, facilitating the intuitive presentation and efficient result sharing. In addition, the visualization module supports dynamic updates. When the input data changes, the system can generate new estimation results in real time.

Regarding model integration, the system builds a real-time data processing platform to achieve seamless integration between the engineering management system and the estimation model. The input data is standardized during integration to ensure that the construction drawings and budget data are consistent with the model input format, thereby ensuring data compatibility. The system uses a Redis caching mechanism and batch processing strategy to optimize response speed while providing real-time feedback functions and automatically generating decision reports to assist project managers in optimizing budget and resource allocation based on the latest estimation results.

3. Experimental Design and Results Analysis

3.1. Experimental Design

The experiments in this study use the data set established in the previous paper, and the experiments are completed on high-performance computers. The software environment is built on the Ubuntu 20.04 operating system, mainly using the Python 3.9 programming language and the TensorFlow 2.10 deep learning framework for model development and training and using tools such as Matplotlib, Seaborn, and D3.js for visualization analysis of the results. This study selects MSE, MAE, R^2, and MAPE as evaluation indicators to evaluate the performance of the construction project cost estimation model. These indicators comprehensively evaluate the model performance from different dimensions, such as the size, direction, and explanatory power of the prediction error, to ensure the objectivity and reliability of the experimental results.

This study conducts a comparative analysis with four typical construction project cost estimation models to verify the superiority of the model constructed in this paper. These control models include the traditional linear regression (LR) model, support vector regression (SVR) model, random forest regression (RFR) model, and gradient boosting decision tree (GBDT) model. As a classic statistical learning method, the linear regression model is mainly used to analyze simple linear relationships. Support vector regression expands nonlinear fitting capabilities through kernel functions. Random forest regression achieves ensemble learning effects by constructing multiple decision trees. Gradient boosting decision tree improves prediction capabilities by gradually optimizing the loss function. These models represent different method systems of traditional statistical learning and machine learning. In the specific experiment, this study randomly selects 5 engineering projects from the data set and uses the experimental group and control group models for cost estimation. The experiment is repeated 10 times. After 10 rounds of estimation, the MSE, MAE, R^2, and MAPE of each model in each round of estimation are calculated.

3.2. Experimental Results and Analysis

Figure 4 shows the MSE of each group in 10 rounds of cost estimation after the experiment:

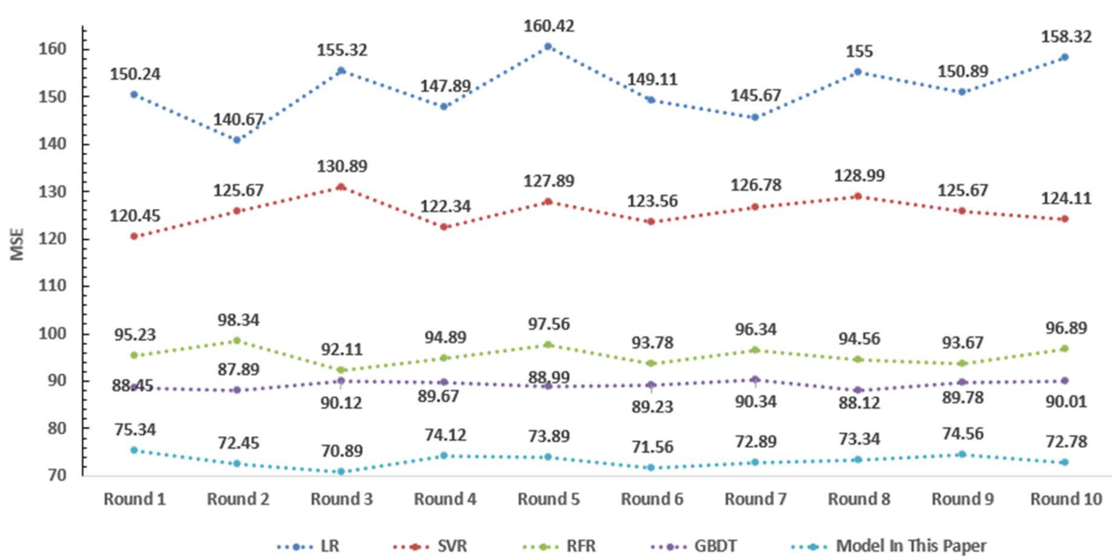

Fig. 4. MSE values of cost estimation of each model

In Figure 4, the performance of the proposed model in construction project cost estimation is better than the other four control models. Specifically, the average MSE value of the traditional LR model is 151.35. The high error indicates that the model has certain limitations in estimating construction project cost. The performance of the SVR model improves slightly, with an average MSE value of 125.64, but it is still higher than the model proposed in this paper. The RFR and GBDT models perform relatively close, at 95.34 and 89.26, respectively. Although these models use ensemble learning methods to improve prediction capabilities, they still fail to surpass the model in this paper.

In contrast, the construction cost estimation model proposed in this paper, which combines subtractive clustering, self-learning mechanism, and CNN, shows better prediction performance, with an average MSE of only 73.18. This result indicates that the model in this paper not only effectively reduces the prediction error but also improves the accuracy and reliability of the estimation. In all 10 rounds of estimation, the MSE value of the model in this paper is generally lower than that of other models, which shows that the application of the model in this paper in the construction cost estimation task has significant advantages and can provide more precise budget support for engineering projects.

In terms of MAE, Table 1 shows the performance of each model:

Table 1. MAE values of estimation of each model

Model	LR	SVR	RFR	GBDT	Model in this paper
Round 1	18.45	15.23	12.56	11.89	9.23
Round 2	16.78	14.89	13.11	11.56	8.45
Round 3	19.23	16.34	12.89	12.34	8.12
Round 4	17.56	15.11	12.23	11.89	8.78
Round 5	20.12	17.23	13.45	12.12	8.56
Round 6	18.34	15.67	12.67	11.45	7.89
Round 7	17.89	16.01	13.12	12.01	8.12
Round 8	18.34	16.45	12.34	11.23	7.87
Round 9	18.78	15.78	12.56	11.67	8.23
Round 10	19.23	16.12	13.01	12.23	8.01

According to the MAE values of each model in 10 rounds of construction project cost estimation in Table 1, it can be found that the model constructed in this paper performs best. Specifically, the MAE value of the model in this paper is always lower than that of other control models, showing its advantage in prediction precision. For example, in estimation round 1, the MAE of the model in this paper is 9.23, while that of other models such as LR is 18.45; that of SVR is 15.23; that of RFR is 12.56; that of GBDT is 11.89, with significant differences. In other rounds of estimation, the MAE value of the model in this paper also remains at the lowest level, with the highest being 8.78, which is significantly lower than the maximum value of other models, showing its stability and reliability in the prediction of various construction projects. In particular, in estimation rounds 6, 8, and 10, the MAE values of the model in this paper are 7.89, 7.87, and 8.01, respectively. This result is better than other models, further proving the model's excellent performance in construction project cost estimation.

In Figure 5, there are significant differences in the performance of each model's coefficient of determination R^2 in the 10 rounds of construction project cost estimation. As the most basic estimation model, LR has an average R^2 of 0.8267 in 10 rounds of estimation, which has the worst performance among all models, indicating that its prediction ability for complex construction projects is limited. SVR performs slightly better than LR, with an average R^2 of 0.8517, but fails to improve the estimation precision significantly. RFR and GBDT, as two ensemble learning models, perform strongly in nonlinear feature fitting, with average R^2 of 0.9118 and 0.9237, respectively, showing better robustness and prediction ability.

In contrast, the construction project cost estimation model constructed in this paper shows the best prediction precision in all experiments, and the average R^2 of 10 estimations reaches 0.9477, which is significantly higher than other models. This is due to the synergy between the subtractive clustering, self-learning mechanism, and CNN module in this model, which can effectively eliminate redundant information, optimize the feature space, and improve the adaptability to complex construction projects through dynamic feedback. Especially in the estimations round 2 and round 10, the R^2 of this model reaches 0.952 and 0.951, respectively, further verifying its strong prediction performance under different types of construction projects. Therefore, this model has obvious advantages in the construction project cost estimation task and is an efficient tool for solving complex engineering cost problems.

Figure 5 presents the performance of each model on R^2:

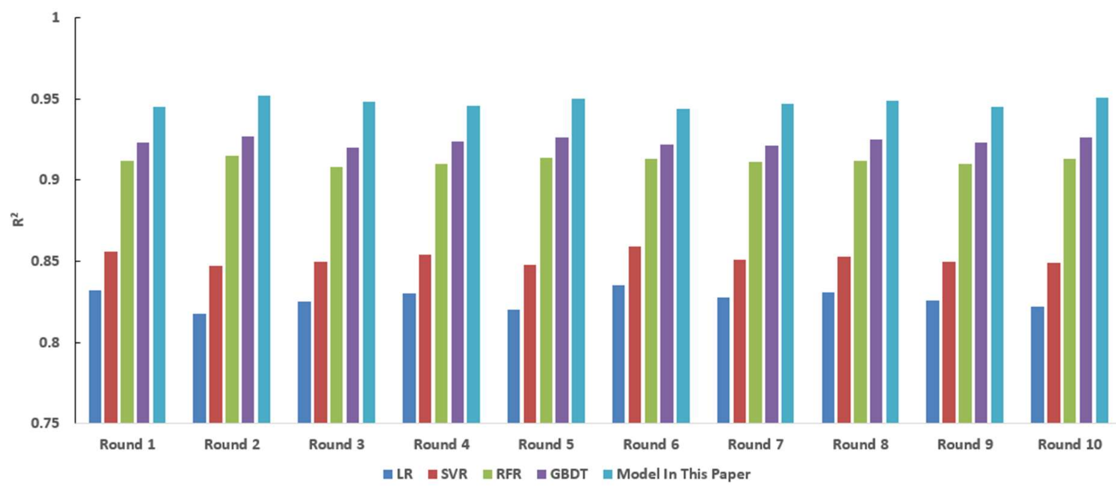

Fig. 5. Results of each model on R^2

Finally, the MAPE of the 10-round estimation results of each group of models is listed in Table 2:

Table 2. MAPE results of 10 rounds of cost estimation of each model

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10
LR (%)	12.45	11.78	12.23	12.01	11.89	12.34	12.12	12.45	12.34	12.23
SVR (%)	10.34	9.89	10.12	10.45	9.78	10.56	10.23	10.67	10.12	10.45
RFR (%)	8.23	8.11	8.45	8.78	8.56	8.67	8.89	8.34	8.56	8.67
GBDT (%)	7.89	7.56	7.34	7.67	7.23	7.12	7.45	7.34	7.67	7.89
Model in this paper (%)	5.67	5.45	5.12	5.34	5.56	5.23	5.12	5.34	5.45	5.01

The data in Table 2 shows that in the 10 rounds of construction project cost estimation, the MAPE value of the LR model is relatively high, with an average of 12.18%, indicating that its precision in cost estimation is relatively low. The SVR model improves the estimation precision to a certain extent, and its average MAPE value drops to 10.26%. The RFR model is further optimized, with a MAPE value of 8.53%, showing high reliability in most cases. The GBDT model performs better than the models above, with a MAPE value reduced to 7.52%, demonstrating strong adaptability and precision. The model constructed in this paper performs best in the estimation results, with an average MAPE value of only 5.33%, which is much lower than other models. This shows that the model in this paper has significant advantages in processing complex and diversified construction project cost data, and its dynamic feedback adjustment and feature optimization capabilities can significantly improve the estimation precision, providing a more precise and robust solution for construction project cost prediction.

4. Conclusion

The construction cost estimation model in this paper shows significant advantages in automatic feature extraction, dynamic optimization, and multimodal data processing. In particular, after combining subtractive clustering, self-learning mechanism, and CNN module, the model achieves efficient feature selection and dynamic parameter adjustment, significantly improving the estimation precision. Experimental results show that the method in this paper shows the best results in multiple construction cost estimations. Its excellent prediction performance proves the reliability and robustness of the model. In addition, the model performs outstandingly in processing multimodal data and can fully mine the potential features in images, texts, and numerical data, ensuring the broad applicability of the model in complex engineering environments.

Despite this, the model still has certain limitations. First, the model's computational efficiency and adaptability may be affected when faced with larger data sets or more complex scenarios. Second, the current model design still relies mainly on integrating existing modules and lacks the application of more cutting-edge deep learning technologies. The model's feature expression capabilities can be further improved in future research by applying advanced deep learning technologies such as Transformer. In addition, in terms of module design, the feedback efficiency of the self-learning mechanism can be further optimized to better cope with dynamically changing engineering scenarios, thereby providing a more comprehensive and efficient solution for construction project cost estimation.

References

[1] Aljohani, A., Ahiaga-Dagbui, D., & Moore, D. (2017). Construction projects cost overrun: What does the literature tell us?. International Journal of Innovation, Management and Technology, 8(2), 137.

[2] Rauzana, A. (2018). Uncertainty variables on cost estimation in project construction. IOSR Journal of Business and Management, 20(1), 80-86.

[3] Durdyev, S. (2021). Review of construction journals on causes of project cost overruns. Engineering, Construction and Architectural Management, 28(4), 1241-1260.

[4] Ibrahim, A. H., & Elshwadfy, L. M. (2021). Factors affecting the accuracy of construction project cost estimation in Egypt. Jordan Journal of Civil Engineering, 15(3).

[5] Sayed, M., Abdel-Hamid, M., & El-Dash, K. (2023). Improving cost estimation in construction projects. International Journal of Construction Management, 23(1), 135-143.

[6] Swei, O., Gregory, J., & Kirchain, R. (2017). Construction cost estimation: A parametric approach for better estimates of expected cost and variation. Transportation Research Part B: Methodological, 101, 295-305.

[7] Elmousalami, H. H. (2020). Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. Journal of Construction Engineering and Management, 146(1), 03119008.

[8] Awosina, A., Ndihokubwayo, R., & Fapohunda, J. (2018). Effects of inaccurate cost estimate on construction project stakeholders. Journal of Construction Project Management and Innovation, 8(2), 1886-1904.

[9] Ahn, J., Ji, S. H., Ahn, S. J., Park, M., Lee, H. S., Kwon, N., ... & Kim, Y. (2020). Performance evaluation of normalization-based CBR models for improving construction cost estimation. Automation in Construction, 119, 103329.

[10] Jung, S., Pyeon, J. H., Lee, H. S., Park, M., Yoon, I., & Rho, J. (2020). Construction cost estimation using a case-based reasoning hybrid genetic algorithm based on local search method. Sustainability, 12(19), 7920.

[11] Sridarran, P., Keraminiyage, K., & Herszon, L. (2017). Improving the cost estimates of complex projects in the project-based industries. Built Environment Project and Asset Management, 7(2), 173-184.

[12] Xiao, X., Skitmore, M., Yao, W., & Ali, Y. (2023). Improving robustness of case-based reasoning for early-stage construction cost estimation. Automation in Construction, 151, 104777.

[13] Abd El-Karim, M. S. B. A., Mosa El Nawawy, O. A., & Abdel-Alim, A. M. (2017). Identification and assessment of risk factors affecting construction projects. HBRC journal, 13(2), 202-216.

[14] Fan, M., & Sharma, A. (2021). Design and implementation of construction cost prediction model based on svm and lssvm in industries 4.0. International journal of intelligent computing and cybernetics, 14(2), 145-157.

[15] Deng, J., & Jian, W. (2022). Estimating construction project duration and costs upon completion using Monte Carlo simulations and improved earned value management. Buildings, 12(12), 2173.

[16] Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. SN Applied Sciences, 2(10), 1703.

[17] Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2022). An artificial neural network approach for cost estimation of engineering services. International journal of construction management, 22(7), 1274-1287.

[18] Fazeli, A., Dashti, M. S., Jalaei, F., & Khanzadi, M. (2021). An integrated BIM-based approach for cost estimation in construction projects. Engineering, Construction and Architectural Management, 28(9), 2828-2854.

[19] Ji, S. H., Ahn, J., Lee, H. S., & Han, K. (2019). Cost estimation model using modified parameters for construction projects. Advances in Civil Engineering, 2019(1), 8290935.

[20] Tijanić, K., Car-Pušić, D., & Šperac, M. (2020). Cost estimation in road construction using artificial neural network. Neural Computing and Applications, 32(13), 9343-9355.

[21] Rafiei, M. H., & Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes. Journal of construction engineering and management, 144(12), 04018106.

[22] Xue, X., Jia, Y., & Tang, Y. (2020). Expressway project cost estimation with a convolutional neural network model. IEEE Access, 8, 217848-217866.

[23] Zhang, H., Wang, W., Zhang, S., Huang, B., Zhang, Y., Wang, M., ... & Wang, Z. (2022). A novel method based on a convolutional graph neural network for manufacturing cost estimation. Journal of Manufacturing Systems, 65, 837-852.

[24] Jiang, Q. (2020). Estimation of construction project building cost by back-propagation neural network. Journal of Engineering, Design and Technology, 18(3), 601-609.

[25] Du, Z., & Li, B. (2017, May). Construction project cost estimation based on improved BP Neural Network. In 2017 International Conference on Smart Grid and Electrical Automation (ICSGEA) (pp. 223-226). IEEE.

[26] Xu, X., Peng, L., Ji, Z., Zheng, S., Tian, Z., & Geng, S. (2021). Research on substation project cost prediction based on sparrow search algorithm optimized BP neural network. Sustainability, 13(24), 13746.

[27] Xin, Z. (2022, December). Research on Engineering Cost Management of Construction Project Based on BIM Technology and BP Neural Network. In 2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT) (pp. 1-5). IEEE.

[28] Wang, X. (2017). Application of fuzzy math in cost estimation of construction project. Journal of Discrete Mathematical Sciences and Cryptography, 20(4), 805-816.

[29] Habibi, F., Birgani, O., Koppelaar, H., & Radenović, S. (2018). Using fuzzy logic to improve the project time and cost estimation based on Project Evaluation and Review Technique (PERT). Journal of Project Management, 3(4), 183-196.

[30] Beilin, I. L. (2017). Economic-Mathematical Modeling of the Cost of Innovation Project Based on Discrete and Continuous Fuzzy Numbers. Journal of Economic & Management Perspectives, 11(3), 1490-1498.

[31] Obianyo, J. I., Okey, O. E., & Alaneme, G. U. (2022). Assessment of cost overrun factors in construction projects in Nigeria using fuzzy logic. Innovative Infrastructure Solutions, 7(5), 304.

[32] Xu, Y., & Cao, S. (2023). Building Engineering Cost Prediction Model Based on TSNE and Improved Grey Correlation Algorithm. Procedia Computer Science, 228, 957-965.

[33] Nadafi, S., Moosavirad, S. H., & Ariafar, S. (2019). Predicting the project time and costs using EVM based on gray numbers. Engineering, Construction and Architectural Management, 26(9), 2107-2119.

[34] Xie, S., & Fang, J. (2018). Prediction of construction cost index based on multi variable grey neural network model. International Journal of Information Systems and Change Management, 10(3), 209-226.

[35] Jia, H. (2022). Construction of Construction Project Budget Estimation Model Based on Grey Markov Chain. Adv. Comput. Signals Syst, 6, 3-10.