The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach

ChoongWan Koo, TaeHoon Hong, ChangTaeK Hyun

Department of Architectural Engineering, Yonsei University, Seoul, Republic of Korea

Department of Architectural Engineering, University of Seoul, Seoul, Republic of Korea

Keywords:
Construction costs
Housing
Regression analysis
Neural networks
Optimization

ABSTRACT

Decision-making in the early stages of a construction project will have a significant impact on the project. Limited and uncertain information, however, makes it difficult to accurately predict construction costs. To solve this problem, this study developed the advanced case-based reasoning (CBR) model with 101 cases of multi-family housing projects.

The advanced CBR model was developed to integrate the advantages of prediction methodologies such as CBR, multiple regression analysis (MRA), and artificial neural networks (ANN), and the optimization process using a genetic algorithm. This study defined four optimization parameters, as follows: (i) the minimum criterion for scoring the attribute similarity, (ii) the range of attribute weight, (iii) the range of case selection and (iv) the tolerance range of cross range between MRA and ANN. Since the system was developed using the Microsoft-Excel-based Visual Basic Application (VBA) for ease of use, it is expected that the model supports the stakeholders in charge of predicting and managing a construction cost in the early stages of a construction project to get more accurate result from historical cases as a reference.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Decision-making in the early stages of a construction project will have a great effect on the project. As a project is implemented, information on it becomes more specific, which makes it more accurate to make decisions such as predicting construction costs. The time and efforts involved in the project also increase, however, and the level of the project’s effectiveness goes down (Construction Industry Institute (CII), 1998; Michael, 1990).

The construction industry has features that are in stark contrast to those of the manufacturing industry, in which products are produced in particular sites based on orders with certain designs (Koo, Hong, Hyun, Park, & Seo, 2010). The features of the construction industry make it more difficult to predict and plan construction projects in their early stages. Therefore, whether or not a certain company or person has an experience similar to particular projects is a critical factor in evaluating the capacity of that company or person.

In case such essential information on the project is not sufficient at its early stages, it is very important to quickly and accurately find particular information from a wealth of historical data. Especially, critical management indices such as cost and schedule must be systematically stored and used. Therefore, it is required to develop the model that is capable of predicting the construction costs in terms of establishing a reasonable decision-making process.

The model that was developed in this study can be used in all types, sizes, and locations of projects. However, this study was designed to find the best solution from historical data in the early stages of a construction projects, thus repetitive projects are more suitable for the model. The project characteristics and cost data from 101 multi-family housing projects that were completed between 2000 and 2005 were used to develop the model in this study (Korea Institute of Construction Technology (KICT), 2007; Korea National Housing Corporation (KNHC), 2005). Since very restricted information was available in the early stages, some information that could be assumed was used to develop the model.

Previous studies have used various prediction methodologies to yield more accurate and reasonable results. Koo et al. (2010) found that such methodologies have distinct characteristics in terms of their applied fields, analysis data, system establishment methods, types of results, and levels of model optimization. Case-based reasoning (CBR) has characteristics that are similar to humans’ heuristic approach, in which decisions are made based on experience (Watson, 1997). Multiple regression analysis (MRA) arrives at the results through a statistical analysis, but its results are too linear to be used in a standardized model (Duverlie & Castelain, 1999; Lowe, Emsley, & Harding, 2006; Phaobunjong, 2002). Artificial neural networks (ANN) is clearly superior to the other methods with...
regard to accuracy, but it has a black box that cannot explain the structure of the model (Attalla & Hegazy, 2003; Hegazy & Ayed, 1998; Rifat, 2004). Genetic algorithm (GA) is a search technique that finds exact and approximate solutions through the repeat process, it is relatively simple to implement (Dogan, Arditi, & Gunaydin, 2006).

In the study conducted by Koo, Hong, Hyun, and Koo (2009), Koo et al. (2010), an integrated model based on several methodologies was developed in which CBR was mainly used. It was designed to coincide with the current practical process. Thus, an integrated model based on CBR approach was used in this study as well, where an engine for filtering the predicted value was applied to improve the prediction accuracy, which is different from the previous study (Koo et al., 2009, 2010). Therefore, "the advanced case-based reasoning approach" was used to express the concept of the model developed in this study.

The model developed in this study was divided into three parts. The first two parts were based on the CBR process, and the third part was based on the GA process. In the first part (Stage 1), the case similarity that consisted of both the attribute similarity and the attribute weight was calculated as a typical CBR process. In the second part (Stage 2), the predicted value as a target variable was selected and filtered based on certain criteria which were deduced from both MRA and ANN. In the third part (Stage 3), GA was used for optimization process, in which some variables that have an effect on the target variable (i.e., the prediction accuracy) were established as the optimization parameters. During the optimization process, the GA finds the optimal value of these parameters within certain ranges. Finally, the system was established on the basis of the process mentioned above using the Microsoft-Excel-based Visual Basic Application (VBA) for ease of use.

2. Process of development of the construction cost prediction model

As shown in Fig. 1, in this study, a model for predicting construction costs was developed using the advanced CBR approach based on the characteristics of multi-family housing projects. In the CBR algorithm, a reasonable environment for selecting the most similar cases could be varied according to internal factors such as the type of attribute scale, the methods of calculating the attribute similarity, the attribute weight, and the case similarity. Those factors depended on the viewpoint of the researcher (Mark, Simoudis, & Hinkle, 1996; Watson, 1997).

In the development process of the model, it was shown that several factors are very essential in determining the prediction capacity of the model. In Stage 1, two optimization parameters were applied to calculate the case similarity: (1) the minimum criterion for scoring the attribute similarity (MCAS) and (2) the range of the attribute weight (RAW). In Stage 2, another two optimization parameters were applied: (3) the range of the case selection (RCS) and (4) the tolerance range of the cross-range between MRA and ANN (TRCRMA), which were related to the improvement of the prediction capacity. In other words, the predicted value, as a target variable, was selected and filtered based on two optimization parameters ((3) and (4)). In previous researches (Koo et al., 2009, 2010), when both the case selection criterion (RCS) and the tolerance range of the cross-range between MRA and ANN (TRCRMA) were used, the prediction accuracy of the model was increased. In this study, the predicted value was selected and filtered based on both optimization parameters ((3) and (4)) for ease of use.

Fig. 1. Process of development of the model through advanced CBR approach.
optimization parameters (1) and (2) were applied to the development of the CBR model, it was proven that the effectiveness of the model was more improved. In this study, optimization parameters (3) and (4) were first adopted, which was appeared to be very important and critical factors. This study set up such factors as the optimization parameters, which were used to develop the optimization process for the improvement of the prediction capacity in Stage 3.

In Stage 3, the optimization process was completed using GA to improve the prediction capacity of the advanced CBR model. First, MCAS, RAW, RCS, and TRCRMA were set at “Adjustable” in the software program “Evolver”, and the optimization process was conducted within 0–100% to calculate the attribute similarity, deduce the attribute weight, select the cases according to the rank of the case similarity, and filter the predicted value. Second, the number of prediction cases (NPC) was set at “Constraint” to exclude the outlier cases, due to which this study was able to control the standard deviation of the prediction accuracy.

Fig. 1 shows the process of development of the advanced CBR model. The process generally consisted of three parts: (i) the process of development of the model from Stages 1–3; (ii) the equations used in each step of the development process; and (iii) the optimization process using GA to improve the prediction capacity. More detailed information is explained in the following section with a representative case (Case 1).

2.1. Stage 1: process of calculation of the case similarity

In Stage 1, the typical process of the CBR was established, which means the reasonable environment for the calculation of the attribute similarity, the attribute weight, and the case similarity. The nearest-neighbor retrieval method was used to calculate the attribute similarity with MCAS as the optimization parameter. GA was used to deduce the attribute weight.

2.1.1. Step 1-1: organizing and formatting of the data

In Step 1-1, the case-base that became the basis for the more effective formulation of a reasonable environment for the CBR model was established. The case-base consisted of the case #, case name, input attributes, and output attributes. The scale of the attribute was defined against all the attributes shown in Table 1, and was linked to the method of calculating the attribute similarity in Step 1-2.

2.1.2. Step 1-2: calculation of the attribute similarity

In Step 1-2, the attribute similarity was calculated. If an attribute was defined as nominal scale, when the value of the attribute is same, the score of the attribute similarity was rated as 1, otherwise 0. If an attribute was defined as either interval or ratio scale, its attribute similarity was scored with Eq. (1), using the nearest-neighbor retrieval method, only when the score of the attribute

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Optimized value of the optimization parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Optimization parameters</td>
<td>(2) No.</td>
</tr>
<tr>
<td>MCAS</td>
<td>–</td>
</tr>
<tr>
<td>X1</td>
<td>Construction duration</td>
</tr>
<tr>
<td>X2</td>
<td>Delivery method-1</td>
</tr>
<tr>
<td>X3</td>
<td>Delivery method-2</td>
</tr>
<tr>
<td>X4</td>
<td>Type of multi-family housing-1</td>
</tr>
<tr>
<td>X5</td>
<td>No. of households-1</td>
</tr>
<tr>
<td>X6</td>
<td>Type of multi-family housing-2</td>
</tr>
<tr>
<td>X7</td>
<td>No. of households-2</td>
</tr>
<tr>
<td>X8</td>
<td>Type of multi-family housing-3</td>
</tr>
<tr>
<td>X9</td>
<td>No. of households-3</td>
</tr>
<tr>
<td>RAW</td>
<td>–</td>
</tr>
<tr>
<td>X10</td>
<td>Site location-1</td>
</tr>
<tr>
<td>X11</td>
<td>Site location-2</td>
</tr>
<tr>
<td>X12</td>
<td>Site location-3</td>
</tr>
<tr>
<td>X13</td>
<td>Site location-4</td>
</tr>
<tr>
<td>X14</td>
<td>Site location-5</td>
</tr>
<tr>
<td>X15</td>
<td>Site location-6</td>
</tr>
<tr>
<td>X16</td>
<td>Non-working days</td>
</tr>
<tr>
<td>X17</td>
<td>Total floor area</td>
</tr>
<tr>
<td>X18</td>
<td>No. of stories above the ground</td>
</tr>
<tr>
<td>X19</td>
<td>No. of stories below the ground</td>
</tr>
<tr>
<td>X20</td>
<td>Size of household</td>
</tr>
<tr>
<td>X21</td>
<td>Land ratio</td>
</tr>
<tr>
<td>RCS</td>
<td>–</td>
</tr>
<tr>
<td>TRCRMA</td>
<td>–</td>
</tr>
</tbody>
</table>

* MCAS = minimum criterion for scoring the attribute similarity.
** RAW = range of attribute weight.
*** RCS = range of the case selection.
**** TRCRMA = tolerance range of the cross range between MBA and ANN, (1) ratio scale: the scale that defines the attributes distinguished by quantifiable values or by ratio, (2) nominal scale: the scale that defines attributes such as objects or class distinguished by name, (3) interval scale: the scale that defines the attributes distinguished by level.
similarity was more than MCAS; otherwise, 0. MCAS was added as the optimization parameter in Stage 3.

\[
f_{AS}(x) = \begin{cases} 
100 - \left( \frac{AV_{Test,Case} \times AV_{Retrieved,Case}}{AV_{Retrieved,Case}} \times 100 \right) & \text{if } f_{AS}(x) \geq MCAS, \\
0 & \text{if } f_{AS}(x) < MCAS.
\end{cases}
\] (1)

where, \( f_{AS} \) is the function of the attribute similarity, \( AV_{Test,Case} \) is the attribute value of the Test_Case, \( AV_{Retrieved,Case} \) is the attribute value of the Retrieved_Case, and MCAS is the minimum score for scoring the attribute similarity. For example, in the case of the total floor area defined as a ratio scale, its attribute similarity was calculated as follows: \( 0.96238813 \times 100 - \left( 0.49208957 \times 0.51059798 \right) / 0.49208957 \times 100 = 74.4008423\% \) when the standardized value \( 0.49208957 \) in Case 1 and \( 0.51059798 \) in Case 14 were applied to Eq. (1). Since MCAS was set at 74.4008423\% through the optimization process using GA in Stage 3 (refer to Table 1), the score for the attribute similarity became valid. Had MCAS been set at 97\%, however, it would not have been recognized, and a score of 0 was thus given.

2.1.3. Step 1-3: calculation of the attribute weight

In Step 1-3, the attribute weight was calculated. In this study, GA was used to derive the attribute weight. In a previous research (Koo et al., 2010), it was found that GA was the best method of calculating the attribute weight. Feature counting, however, was used for the control group. A detailed description of the two methodologies was as follows.

- Feature counting (FC): this method applies 1 as a weight to all the attributes, which is based on the understanding that there is no need to apply to them a higher value than 1.
- GA: this method optimizes the value of the attribute weight with the target based on the prediction accuracy, where the attribute weights could be changed within a range using GA.

RAW was applied to the optimization process in Stage 3 to find the best value for the attribute weight. The optimized value of the attribute weight was deduced through the optimization process using GA in Stage 3 (refer to the fifth column [(5) optimized value] in Table 1).

2.1.4. Step 1-4: calculation of the case similarity

In Step 1-4, the case similarity was calculated. The method of calculating the attribute weight was introduced in Step 1-3. Eq. (1) shows the method of calculating the attribute similarity in Step 1-2. By multiplying these two values, the weighted-attribute similarity was derived. The accumulated sum of such value based on the attribute (attribute weight + attribute similarity) was divided by the accumulated sum of the attribute weight to calculate the case similarity score. The case similarity score was calculated using Eq. (2).

\[
f_{CS}(x) = \frac{\sum_{i=1}^{n} (f_{AS_i} \times f_{AW_i})}{\sum_{i=1}^{n} (f_{AW_i})},
\] (2)

where, \( n \) is the number of attributes, \( f_{CS} \) is the function of the case similarity, \( f_{AS} \) is the function of the attribute similarity, and \( f_{AW} \) is the function of the attribute weight. For example, the procedure for calculating the case similarity of Cases 1 and 14 is as follows. The attribute similarity of the total floor area was found to have been 0.96238813 (96.23%\%) using Eq. (1). As shown in the fifth column of Table 1 [(5) optimized value], the attribute weight of the total floor area was deduced as 0.81313608 through the optimization process using GA. The multiplication of these two values yielded 0.78255251 (0.96238813 \times 0.81313608), which is the weighted attribute similarity. By dividing the accumulated sum (4.58297081) of the weighted attribute similarity by the accumulated sum (11.08346036) of the attribute weight, the case similarity was calculated as 0.41349639 (4.58297081 / 11.08346036) using Eq. (2).

2.2. Stage 2: process of improvement of the prediction capacity

In Stage 2, the advanced CBR approach was introduced as a new concept for improving the prediction capacity. In previous studies (Dogan et al., 2006; Koo et al., 2009, 2010), the cases were ranked from those with higher case similarities to those with lower case similarities using a formula similar to Eq. (2), after which the cases were retrieved only by rank according to the case similarity score. In this study, however, the predicted value, as a target variable, was selected and filtered based on the following criteria: (i) the range of the case selection (RCS) and (ii) the cross-range between MRA and ANN to which TRCRMA was applied (CRMA*) (refer to Table 2 and Fig. 2). Detail information on RCS and CRMA* were described in Steps 2-1 and 2-2.

2.2.1. Step 2-1: selection of the cases within specific ranges

In Step 2-1, the cases ranked according to their case similarities calculated in Step 1-4 of Stage 1 and specific ranges for selecting the cases were found during the optimization process. According to the concept of the CBR approach, the case with the highest case similarity score among the case-base may be considered to have the most similar project characteristics compared to the Test_Case. In this study, the advanced CBR model was developed in compliance with this basic concept as well.

Whereas the number of cases that would be finally selected was predefined prior to the optimization process in previous studies (Kim, Kim, & Kang, 2004), the advanced CBR model developed in this study was designed to decide on the number of cases that would be finally selected within specific ranges through the optimization process. RCS was applied to the optimization process as the optimization parameter in Stage 3.

For example, since RCS was set at 19.0329446\% through the optimization process using GA in Stage 3 (refer to Table 1) it may be considered that the cases ranked from 1 to 19 among the case-base according to the case similarity score were selected and used as references.

2.2.2. Step 2-2: filtering of the predicted value

In Step 2-2, the cross-range between the predicted value of MRA and the ANN model (CRMA) was produced to filter the predicted value. The predicted value was calculated within the range of the minimum, the most likely, and the maximum, using the standard error rate. This range of the predicted value was presented differently by model (i.e., MRA and ANN). CRMA as a filtering engine was deducted through a synthesis analysis of both MRA and the ANN model. TRCRMA was added as the optimization parameter to give CRMA a tolerance range in Stage 3. Fig. 2 shows a graph with which the concept of CRMA, TRCRMA, and CRMA* can be explained. Its detailed description and figures are as follows.

2.2.2.1. Step 2-2-1: calculation of the range of the predicted value using the MRA model

Eq. (3) was used to calculate the range of the predicted value using the MRA model:

\[
P_{V_{MRA}} \times (1 - (SER_{MRA} / 100)) \leq PR_{MRA} \leq PV_{MRA} \times (1 + (SER_{MRA} / 100)),
\] (3)

where \( PV_{MRA} \) is the predicted value of the MRA model, \( SER_{MRA} \) is the standard error rate of the MRA model, and \( PR_{MRA} \) is the predicted range of the MRA model. For example, the predicted value of Case
1 using the MRA model \(P_{V_{MRA}}\) was deduced as 43,313.38 using the software program SPSS 12.0. The standard error rate of the MRA model \(SER_{MRA}\) was calculated as 15.18875240%. Thus, the predicted range of the MRA model \(PR_{MRA}\) was deduced as 36,734.62 \(\leq P_{V_{MRA}} \leq 49,892.14\) using Eq. (3) (refer to Fig. 2).

2.2.2.2. Step 2-2-2: calculation of the range of the predicted value using the ANN model. Eq. (4) was used to calculate the range of the predicted value in the ANN model:

\[
P_{V_{ANN}} \times (1 - \text{SER}_{ANN}/100) \leq PR_{ANN} \leq P_{V_{ANN}}(1 + \text{SER}_{ANN}/100)
\]

where \(P_{V_{ANN}}\) is the predicted value of the ANN model, \(\text{SER}_{ANN}\) is the standard error rate of the ANN model, and \(PR_{ANN}\) is the predicted range of the ANN model. For example, the predicted value of Case 1 in the ANN model \(P_{V_{ANN}}\) was deduced as 46,939.47 with the software program Neuro Solutions 5. The standard error rate of the ANN model \(\text{SER}_{ANN}\) was calculated as 10.26612982%. Thus, the predicted range of the ANN model \(PR_{ANN}\) was deduced as 42,120.60 \(\leq P_{V_{ANN}} \leq 51,758.34\) using Eq. (4) (refer to Fig. 2).

2.2.2.3. Step 2-2-3: deduction of the cross-range between the predicted values of MRA and ANN model. Eq. (5) was used to deduce the cross-range between the predicted values of MRA and ANN model (CRMA):

\[
\text{Max}(\text{Min}(PR_{MRA})), \text{Min}(P_{V_{ANN}}) \leq \text{CRMA} \leq \text{Min}(\text{Max}(PR_{MRA})), \text{Max}(P_{V_{ANN}})
\]

1 TRCRMA was set at 4.9333423% through the optimization process using GA, which was applied to CRMA to make CRMA*.

1 Filtering range of both CRMA and CRMA*:

1 (1) 36,734.62 \(\leq PR_{MRA} \leq 49,892.14\); (2) 42,120.60 \(\leq PR_{ANN} \leq 51,758.34\); (3) 42,120.60 \(\leq \text{CRMA} \leq 49,892.14\); and (4) 40,042.65 \(\leq \text{CRMA*} \leq 52,353.49\).
2.2.4. Step 2-2-4: deducing CRMA* where TRCRMA is applied to CRMA. Eq. (6) is used to calculate the CRMA* that is the filtering range where TRCRMA was applied to CRMA:

\[
\text{Min}(\text{CRMA}) \times (1 - \text{TRCRMA}/100) \leq \text{CRMA} \\
\leq \text{Max}(\text{CRMA}) \times (1 + \text{TRCRMA}/100),
\]

where CRMA is the cross range between the predicted value of MRA and ANN model. TRCRMA is the tolerance range of CRMA, CRMA* is the filtering range that TRCRMA was applied to CRMA. For example, as shown in step 2-2-3, the CRMA of Case #1 was deduced as 42,120.60 \( \leq \) CRMA \( \leq \) 49,892.14 by Eq. (5). Since TRCRMA was set at 4.9333423% through the optimization process using GA in STAGE 3 (refer to Table 1), CRMA* was deduced as 40,042.65 (= 42,120.60 \times (1 - (4.9333423 \div 100))) \( \leq \) CRMA \( \leq \) 52,353.49 (= 49,892.14 \times (1 + (4.9333423 \div 100))) by Eq. (6) (refer to Fig. 2).

2.2.3. Step 2-3: analysis of the prediction capacity

In Step 2-3, the prediction accuracy was calculated. The Retrieved Case was found from the case-base and compared with the Test Case, the project characteristics of which were determined to have been most similar to those of the Test Case. Moreover, the standard error rate and the prediction accuracy were finally calculated. Eq. (7) was used to calculate the standard error rate:

\[
f_{\text{SER}}(x) = \left| \frac{V_{\text{Test Case}} - PV_{\text{Retrieved Case}}}{V_{\text{Test Case}}} \right| \times 100.
\]

where \( f_{\text{SER}} \) is the function of the standard error rate, \( V_{\text{Test Case}} \) is the value of the Test Case, and \( PV_{\text{Retrieved Case}} \) is the predicted value of the Retrieved Case.

Eq. (8) was used to calculate the prediction accuracy:

\[
f_{\text{PA}}(x) = 100 - f_{\text{SER}}(x),
\]

where \( f_{\text{PA}} \) is the function of the prediction accuracy and \( f_{\text{SER}} \) is the function of the standard error rate. For example, Cases 14, 34, 2, and 53 were selected and filtered using Eqs. (2) and (6) as projects that are similar to Case 1 (refer to Table 2). The construction cost for Cases 1 and 14 were 48,771.26 and 48,768.21, respectively. When these two values were applied to Eq. (7), the standard error rate became 0.0062% (= \( \text{abs}(48,771.26 - 48,768.21) \div 48,771.26 \times 100 \)). When \( f_{\text{SER}} \) was applied to Eq. (8), the prediction accuracy became 99.99%. The same procedure and equations for the calculation of the prediction accuracy were applied to Cases 34, 2, and 53, and the prediction accuracy became 91.46%, 96.63%, and 94.17%, respectively. In conclusion, the average of prediction accuracy on the four cases (14, 34, 2, and 53) was found to have been 95.57% (refer to Table 5).

2.3. Stage 3: process of optimization of the prediction model

Fig. 3 shows the correlation between the case similarity score and the prediction accuracy. As shown in Fig. 3, there was a wide of values for the prediction accuracy, as illustrated by the dotted line. It was also shown that the correlation between the prediction accuracy, as illustrated by the dotted line, and the case similarity, as illustrated by the solid line, was not consistent. For instance, as shown in the part of the figure indicated by a circle, there were cases in which the case similarity was high but the prediction accuracy was extremely low (refer to circle (1) in Fig. 3) or vice versa (refer to circle (2) in Fig. 3). It appears that the correlation between the case similarity and the prediction accuracy is not always proportional.

As mentioned in Stage 1, it was found that MCAS and RAW are the main factors in calculating the case similarity score. In a previous research (Koo et al., 2010), it was proven that these two optimization parameters are very important in developing a construction cost prediction model using the CBR approach. Therefore, MCAS and RAW were applied to the development of the model in this study.

Although these two optimization parameters were applied to the development of the model, it still appears that the correlation between the case similarity and the prediction accuracy is not always proportional. Thus, the predicted value should be selected and filtered (refer to circle (3) in Fig. 3). To solve the problem, another two optimization parameters were introduced in this study, to filter the predicted value. It was shown that RCS and TRCRMA

---

**Fig. 3.** Correlation between the case similarity score and the prediction accuracy (Case 1).
are the critical factors in improving the prediction capacity in Stage 2 (refer to Tables 2 and 3).

Therefore, this study defined such factors as the optimization parameters and established the optimization process using GA. Steps 3-1 to 3-6 show the framework of the optimization process for improving the prediction capacity using the advanced CBR approach.

2.3.1. Step 3-1: the 1st optimization parameter – minimum criteria for scoring the attribute similarity

In the study of Koo et al. (2009, 2010), the prediction accuracy was more improved when MCAS was used to calculate the attribute similarity. Therefore, MCAS was defined as the optimization parameter in this study and optimized using GA based on the range of 0–100%.

2.3.2. Step 3-2: the 2nd optimization parameter – range of the attribute weight

To improve the prediction accuracy, Koo et al. (2009) used various methodologies such as ANN, MRA, and FC to deduce the attribute weight. When ANN was used to deduce the attribute weight, the prediction accuracy was greater than with the other methodologies. In the study of Koo et al. (2010), GA was used to calculate the attribute weight, where the target was based on the prediction accuracy. It was shown that when the attribute weight was optimized by itself using GA, the prediction accuracy was more improved.

Therefore, GA was used to deduce the attribute weight in this study and to optimize the value of the attribute weight by itself. The software “Evolver” was used to conduct the optimization process based on the range of 0.00–1.00.

2.3.3. Step 3-3: the 3rd optimization parameter – range of the case selection

The criteria for case selection used in previous studies (Arditi & Tokdemir, 1999; Koo et al., 2009, 2010) enabled selection of the case with the highest case similarity score or a similarity score of 75 or more. The criteria were not based on any logical rule but merely on subjective judgment. In this study, the number of similar cases that would be finally selected was decided on through the optimization process. RCS was defined as the optimization parameter using GA based on the range of 0–100%.

2.3.4. Step 3-4: the 4th optimization parameter – tolerance range of the cross-range between MRA and ANN

Previous studies did not apply the concept of a filtering mechanism such as CRMA, TRCRMA, or CRMA*. Those parameters were used to improve the prediction capacity in this study. CRMA, however, may be too narrow as a filtering range to be effective, and thus, a tolerance range must be applied to CRMA. That is, the concept of TRCRMA was adopted to improve the effectiveness of the model in this study. TRCRMA was defined as the optimization parameter using GA based on the range of 0–100%.

The concept of TRCRMA is explained in detail using Tables 2 and 3. As shown in Case 1 in Table 2, among the cases selected within the 19th rank via RCS, four cases were filtered with 42,120.60 ≤ CRMA ≤ 49,892.14, which were retrieved using Eq. (5) (refer to (5) predicted value filtered by CRMA in Table 2) and another four cases were filtered with 40,042.65 ≤ CRMA* ≤ 52,353.49, which were retrieved using Eq. (6) (refer to (6) predicted value filtered CRMA* in Table 2). The results filtered by the two filtering mechanism (CRMA and CRMA*) are identical, however, to the gray part (refer to the fifth and sixth columns of Table 2 [(5) predicted value filtered by CRMA and (6) predicted value filtered by CRMA*]), and thus the effect of the application of CRMA* was not shown in Case 1.

Meanwhile, as shown in Table 3, Case 72 clearly shows the effect of the application of CRMA*. Although 15,300.44 ≤ CRMA ≤ 16,386.93, which was the result of Eq. (5), did not filter any case as shown in the fifth column [(5) Predicted Value filtered by CRMA], 14,545.61 ≤ CRMA* ≤ 17,195.36 from Eq. (6) did filter three cases as shown in the sixth column [(6) predicted value filtered by CRMA*]. As such, TRCRMA implemented CRMA* with a sufficient tolerance range to solve the unwanted response caused by the CRMA with a too narrow tolerance range.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Case number</th>
<th>Case similarity score</th>
<th>Prediction accuracy</th>
<th>Predicted value filtered by CRMA</th>
<th>Predicted value filtered by CRMA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
<td>96.17991949</td>
<td>70.08053452</td>
<td>11575.38623</td>
<td>11575.38623</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>94.71121284</td>
<td>73.94187771</td>
<td>12211.17444</td>
<td>12211.17444</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>91.70760918</td>
<td>81.08078006</td>
<td>13352.12572</td>
<td>13352.12572</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>89.29874029</td>
<td>83.42134734</td>
<td>13778.92338</td>
<td>13778.92338</td>
</tr>
<tr>
<td>5</td>
<td>69</td>
<td>82.48743226</td>
<td>86.87114090</td>
<td>18685.79124</td>
<td>18685.79124</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>79.48230709</td>
<td>82.16758852</td>
<td>14570.59411</td>
<td>14570.59411</td>
</tr>
<tr>
<td>7</td>
<td>63</td>
<td>77.85311145</td>
<td>91.4237993</td>
<td>15070.74390</td>
<td>15070.74390</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>74.72900797</td>
<td>98.19483470</td>
<td>11414.22733</td>
<td>11414.22733</td>
</tr>
<tr>
<td>9</td>
<td>84</td>
<td>67.55786276</td>
<td>90.34820270</td>
<td>23066.55270</td>
<td>23066.55270</td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>67.56670899</td>
<td>88.84159989</td>
<td>20012.04174</td>
<td>20012.04174</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
<td>67.08953544</td>
<td>81.46237457</td>
<td>13485.08131</td>
<td>13485.08131</td>
</tr>
<tr>
<td>12</td>
<td>76</td>
<td>65.89479349</td>
<td>90.53570586</td>
<td>23861.54004</td>
<td>23861.54004</td>
</tr>
<tr>
<td>13</td>
<td>68</td>
<td>60.60821215</td>
<td>93.09701028</td>
<td>22279.92436</td>
<td>22279.92436</td>
</tr>
<tr>
<td>14</td>
<td>47</td>
<td>60.37531078</td>
<td>97.33042681</td>
<td>23563.10869</td>
<td>23563.10869</td>
</tr>
<tr>
<td>15</td>
<td>77</td>
<td>60.24020212</td>
<td>91.08020591</td>
<td>24609.36555</td>
<td>24609.36555</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>57.75052784</td>
<td>96.84560101</td>
<td>17097.74557</td>
<td>17097.74557</td>
</tr>
<tr>
<td>17</td>
<td>83</td>
<td>55.96432066</td>
<td>59.02527081</td>
<td>22385.17726</td>
<td>22385.17726</td>
</tr>
<tr>
<td>18</td>
<td>56</td>
<td>55.79534094</td>
<td>71.04307138</td>
<td>11734.37098</td>
<td>11734.37098</td>
</tr>
<tr>
<td>19</td>
<td>44</td>
<td>55.21440202</td>
<td>70.78694335</td>
<td>11692.06564</td>
<td>11692.06564</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>54.42260658</td>
<td>83.42671891</td>
<td>13779.81061</td>
<td>13779.81061</td>
</tr>
</tbody>
</table>

TRCRMA was set at 4.9333423% through the optimization process using GA, which was applied to CRMA to make CRMA*. The cases filtered by CRMA and CRMA* are shaded. Since RCS was set at 19.0329446% through the optimization process using GA, the cases were selected within the 19th rank among the case-base, which is indicated with a dotted line (—).

Filtering range of both CRMA and CRMA*: (1) 15,300.44 ≤ MRA ≤ 20,780.71; (2) 13,355.8 ≤ ANN ≤ 16,386.93; (3) 15,300.44 ≤ CRMA ≤ 16,386.93; and (4) 14,545.61 ≤ CRMA* ≤ 17,195.36.
2.3.5. Step 3-5: constraint – the number of prediction cases

In this study, the prediction accuracy was used as a criterion for evaluating the effectiveness of the model. To make the prediction accuracy consistent, not only should the average of the prediction accuracy be controlled, but the standard deviation of the prediction accuracy should also be considered. According to previous studies (Koo et al., 2009, 2010), even if the average of the prediction accuracy is high, the prediction accuracy of a certain case would still be extremely low. Thus, this study developed the model with the exception of the cases that were detected as outliers. NPC was defined as a “Constraint” to exclude the outlier cases for which GA was used based on the significance level of 5%.

2.3.6. Step 3-6: the optimization process using genetic algorithms

As shown in Fig. 1, the optimization process was performed in this study using GA. (1) In Fig. 1 shows the process for calculating the case similarity explained in Stage 1. Eq. (1) calculates the attribute similarity, where MCAS is included as an optimization parameter. Eq. (2) calculates the case similarity, where RAW is included as an optimization parameter. (2) In Fig. 1 shows the process for improving the prediction capacity explained in Stage 2. Some criteria were applied to the optimization process using GA to select and filter the predicted value, which means RCS and TRCRMA. Eqs. (3)–(6) calculated the CRMA. Eqs. (7) and (8) calculated the prediction accuracy. (3) In Fig. 1 shows the process for optimizing

Table 4
Results of the descriptive analysis by model.

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Attribute weight</th>
<th>No. of cases</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
<th>5th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBR-1 (Koo et al., 2009) including MCAS and TAW as optimization parameters</td>
<td>Feature Counting</td>
<td>99</td>
<td>77.092</td>
<td>17.923</td>
<td>79.357</td>
<td>4.677</td>
<td>99.981</td>
<td>43.12</td>
</tr>
<tr>
<td></td>
<td>MRA(orig.)</td>
<td>99</td>
<td>76.388</td>
<td>21.267</td>
<td>80.181</td>
<td>4.415</td>
<td>99.424</td>
<td>42.46</td>
</tr>
<tr>
<td></td>
<td>MRA(abs)</td>
<td>100</td>
<td>78.235</td>
<td>19.405</td>
<td>82.906</td>
<td>4.415</td>
<td>99.265</td>
<td>48.10</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>99</td>
<td>81.078</td>
<td>18.475</td>
<td>86.702</td>
<td>3.652</td>
<td>99.981</td>
<td>54.90</td>
</tr>
<tr>
<td>CBR-2 (Koo et al., 2010) including MCAS and RAW as optimization parameters</td>
<td>GA</td>
<td>100</td>
<td>83.540</td>
<td>19.920</td>
<td>86.702</td>
<td>4.415</td>
<td>99.981</td>
<td>52.47</td>
</tr>
<tr>
<td>CBR-3 (this study) Including MCAS, RAW, RCS, and TRCRMA as optimization parameters</td>
<td>GA</td>
<td>95</td>
<td>87.428</td>
<td>9.066</td>
<td>89.692</td>
<td>47.264</td>
<td>99.099</td>
<td>74.284</td>
</tr>
</tbody>
</table>

Fig. 4. Graphical user interface of the advanced CBR model.
the prediction model. The optimization parameters explained from Steps 3-1 to 3-5 in Stage 3 were applied to the optimization process using GA to maximize the prediction accuracy. As the optimization process was in progress, the optimized value at a point in time and the average value by that time were calculated. When these two values were the same, the optimization process was considered completely implemented, and the optimized values of the optimization parameters were stored (refer to Table 1).

3. Systemization of the advanced CBR model

3.1. Considerations for the systemization

As mentioned earlier, the development of the advanced CBR model was complicated because it involved many steps. Thus it was necessary to systemize the process of the development of the model. In this study, a Microsoft-Excel-based VBA was used to construct the system. The considerations for the systemization are as follows:

- Simplicity of performance of the prediction tasks;
- Accuracy and logic of using new projects as references;
- Clarity of the final results;
- Ease of adapting the change in the case-base or the input variable; and
- Flexibility of the application to different types of repetitive projects.

3.2. Graphical user interface (GUI)

As shown in Fig. 4, the GUI of the advanced CBR model was designed for users to be able to easily access the model. A detailed description of the system is as follows:

- Part (1): this is the phase in which the system is initialized to perform the prediction process. As the user selects a specific type of project, the system reads the corresponding information (for example, a multi-family housing project) from the case-base stored in the system.
- Part (2): the user defines the information on the types of optimization parameters and the scope of the variations to implement the optimization process. The optimized value by optimization parameters are found using GA as explained in Stage 3, which are set up on the basis of the project information stored in case-base.
- Part (3): the user enters the case information such as the total floor area and selects the scale type by variable (for example, the nominal scale, ratio scale, interval scale, etc.) The internal process of the system calculates the case similarity score. In addition, the predicted values are selected and filtered by RCS and CRMA through the optimization process.
- Part (4): the internal process of the system produces the most similar cases to the case information that user entered. Depending on the optimization result, the similar case can be one or more cases. The final result shows not only the predicted construction cost but also the information related to the similar case. To help the user understand better the final result, the system also produces visual data such as via graphs and tables.

4. Results and discussion

As mentioned earlier, the correlation between the case similarity and the prediction accuracy is not always proportional. Thus, four factors (MCAS, RAW, RCS, and TRCRMA) for improving the
prediction accuracy were established as the optimization parameters in this study. Table 4 shows the results of the descriptive analysis of the prediction accuracy by methodology. As shown in the fourth column [(4) mean] in Table 4, the average value of the prediction accuracy of the model that was developed in this study was 87.428%, where MCAS, RAW, RCS, and TRCRMA were set through the optimization process using GA, which were shown in the fifth column [(5) optimized value] in Table 1.

As shown in the fourth column [(4) mean] in Table 4, the advanced CBR model (CBR-3) that was developed in this study was improved and the prediction accuracy became higher than that of CBR-1 (developed by Koo et al. (2009)) and CBR-2 (developed by Koo et al. (2010)). The average prediction accuracy in the CBR-3 model was 87.43%, which was greater than that of CBR-2 (83.54%) and CBR-1 (81.08%). This means that RCS and TRCRMA, which were newly used as the optimization parameters to develop the model in this study, could be adapted to future researches to improve the prediction accuracy.

As shown in the fifth column [(5) standard deviation] in Table 4, the standard deviation of CBR-3 was lowest (9.066%). This means that the model developed in this study could control the standard deviation of the prediction accuracy and was optimized enough to get consistent prediction accuracy.

In conclusion, as the case-base or project information may be changed, the optimization process of the advanced CBR model could be reactivated to find the optimized value. It was proven that the advanced CBR model could improve the prediction accuracy by itself by finding the optimized values of the optimization parameters using GA. In future studies, the prediction capability of the proposed cost estimation method could be further improved.

5. Validation

As shown in Table 4, the average prediction accuracy of the model developed in this study was 87.428% and the standard deviation was 9.066%. This is the result for verification on all projects of the case-base that was used to develop the model in this study.

In case of case 1 (refer to Table 5), some cases similar to case 1 were retrieved from the case-base. Some cases contained not only the predicted construction cost but also the project characteristics in both the test case and the retrieved case. These results may be used as references in the decision-making process.

As mentioned in Steps 2-2, 3-3, and 3-4, Cases 14, 34, 2, and 53 were selected and filtered as projects similar to Case 1 (refer to Table 2). The average prediction accuracy in the four cases (14, 34, 2, and 53) was appeared as 95.57%.

6. Conclusion

This study developed the advanced CBR model for predicting construction costs based on the characteristics of multi-family housing projects. This model integrated the advantages of (i) prediction methodologies such as CBR, ANN, and MRA with (ii) the optimization process using GA. This study especially defined several optimization parameters such as not only MCAS and RAW, which had been proven in previous studies (Koo et al., 2009, 2010), but also RCS and TRCRMA, which were newly used in this study. This optimization process was completed using GA.

The average prediction accuracy of the advanced CBR model was deduced as 87.428%, where MCAS, RAW, RCS, and TRCRMA were set at the optimized value through the optimization process as shown in the fifth column [(5) optimized value] in Table 1.

The advanced CBR model (CBR-3) was developed to make the prediction capacity better than that in previous studies (Koo et al., 2009, 2010). This means that the optimization parameters such as MCAS, RAW, RCS, and TRCRMA should be adopted and considered in the development of the CBR process.

As mentioned in previous studies (Koo et al., 2009, 2010), the advanced CBR model that was developed in this study is a flexible tool in terms of expansion. In developing a CBR algorithm for different types of projects, the optimization process that was used in this study could be applied to the improvement of the prediction capacity. The optimization process could be reactivated whenever the data are changed. It could not be applied to unique types of projects, however, as they do not have historical data.

The advanced CBR model is a useful tool that can, by itself, optimize the construction cost prediction process for reasonable decision-making. It is expected that this tool will support stakeholders who are in charge of predicting and managing construction costs in the early stages of a construction project. Since the system was developed using Microsoft-Excel-based VBA, the user can correctly, quickly, and easily find useful results by applying this system.

Meanwhile, it was proven that the model developed in this study would be a very effective supporting tool in decision-making process, this study and previous researches were focused on the retrieved process that is just one step of the CBR process. Therefore, additional studies on the revised process, which is another step of the CBR process, need to be conducted to improve the prediction capacity in future studies.

References