Expert Systems with Applications 38 (2011) 8597–8606

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

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

ChoongWan Koo [a], TaeHoon Hong [a,*], ChangTaek Hyun [b]

[a] Department of Architectural Engineering, Yonsei University, Seoul, Republic of Korea
[b] Department of Architectural Engineering, University of Seoul, Seoul, Republic of Korea

ARTICLE INFO

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 constriction 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.

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

* Corresponding author. Tel.: +82 2 2123 5788; fax: +82 2 365 4665.
E-mail address: hong7@yonsei.ac.kr (T. Hong).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.063

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

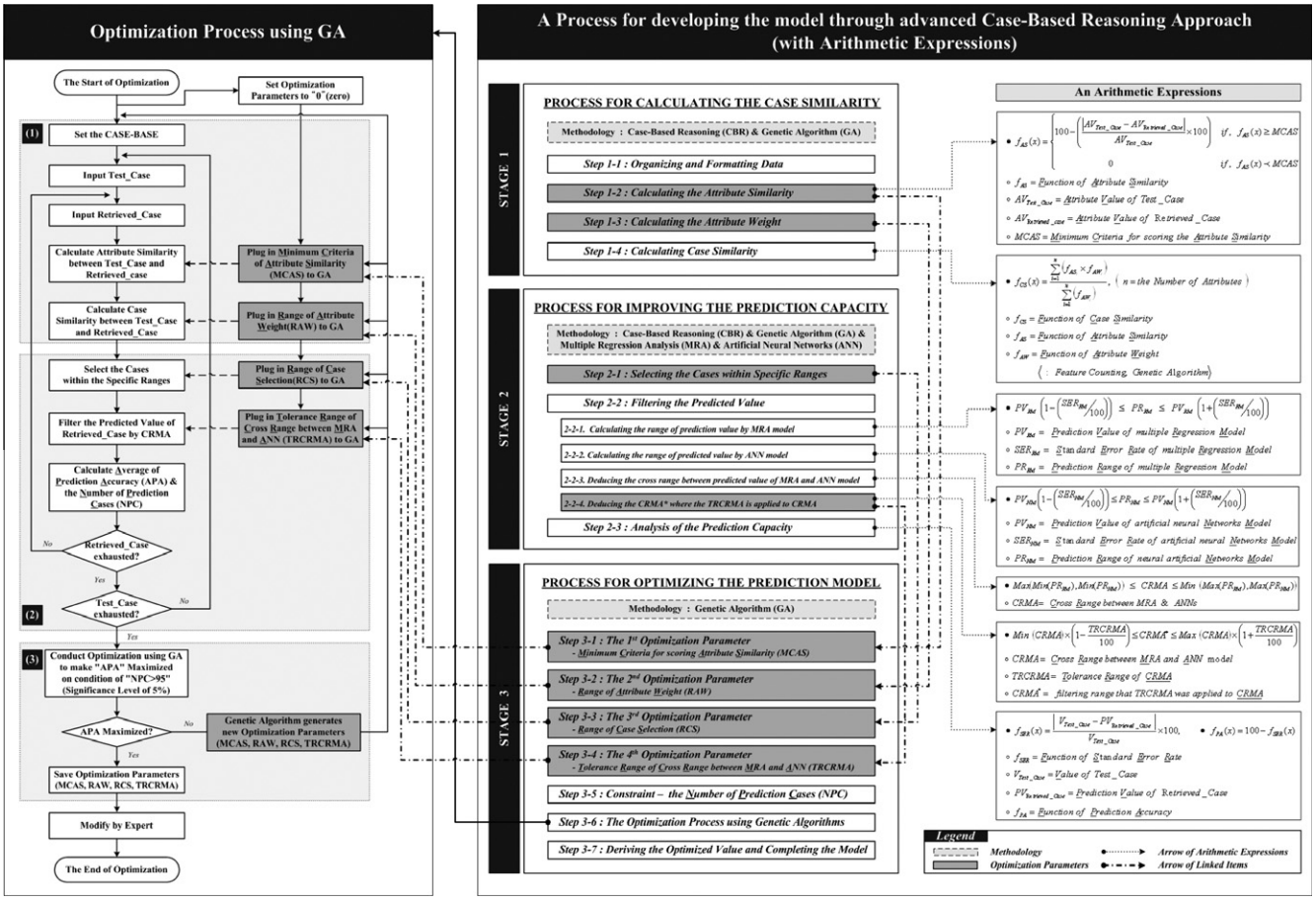

Fig. 1. Process of development of the model through the advanced CBR approach.

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 1
Optimized value of the optimization parameters.

(1) Optimization parameters	(2) No.	(3) Attributes	(4) Type of scale	(5) Optimized value
MCAS*	–	–	–	0.74400842
	X_1	Construction duration	Ratio scale(1)	0.80804276
	X_2	Delivery method-1 (Design-build)	Nominal scale(2)	0.43186254
	X_3	Delivery method-2 (Design-bid-build)	Nominal scale	0.47239523
	X_4	Type of multi-family housing-1 (Public sale)	Nominal scale	0.54179925
	X_5	No. of households-1 (Public sale)	Ratio scale	0.55360367
	X_6	Type of multi-family housing-2 (Public lease)	Nominal scale	0.41571743
	X_7	No. of households-2 (Public lease)	Ratio scale	0.64982608
	X_8	Type of multi-family housing-3 (Permanent rental)	Nominal scale	0.53434482
	X_9	No. of households-3 (Permanent rental)	Ratio scale	0.50040000
RAW**	X_{10}	Site location-1 (District factor-1: Sudo)	Nominal scale	0.28311779
	X_{11}	Site location-2 (District factor-2: Kangwon)	Nominal scale	0.93580884
	X_{12}	Site location-3 (District factor-3: Choongchung)	Nominal scale	0.38034932
	X_{13}	Site location-4 (District factor-4: Junla)	Nominal scale	0.66392902
	X_{14}	Site location-5 (District factor-5: Kyoungsang)	Nominal scale	0.48888315
	X_{15}	Site location-6 (District factor-6: Jeju)	Nominal scale	0.14473379
	X_{16}	Non-working days	Interval scale(3)	0.45370811
	X_{17}	Total floor area	Ratio scale	0.81313608
	X_{18}	No. of stories above the ground	Ratio scale	0.52226251
	X_{19}	No. of stories below the ground	Ratio scale	0.54035077
	X_{20}	Size of household	Ratio scale	0.49207242
	X_{21}	Land ratio	Ratio scale	0.45711680
RCS***	–	–	–	0.19032945
TRCRMA****	–	–	–	0.04933342

* 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 MRA 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.

$$f_{as}(x) = \begin{cases} 100 - \left(\frac{|AV_{Test\_Case} - AV_{Retrieved\_Case}|}{AV_{Test\_Case}} \times 100\right) & \text{if} \quad f_{AS}(x) \geqslant MCAS, \\ 0 & \text{if} \quad f_{AS}(x) < MCAS, \end{cases} \tag{1}$$

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 criterion 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 {96.239% = 100 − [abs (0.49208957 − 0.51059798) ÷ 0.49208957 × 100]} 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).

$$f_{CS}(x) = \frac{\sum_{i=1}^{n}(f_{AS_i} * f_{AW_i})}{\sum_{i=1}^{n}(f_{AW_i})}, \tag{2}$$

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 (96.239%) 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 × 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:

$$PV_{MRA} \times (1 - (SER_{MRA}/100)) \leqslant PR_{MRA} \leqslant PV_{MRA}(1 + (SER_{MRA}/100)), \tag{3}$$

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

Table 2
The similar cases produced by filtering engines (i.e., RCS and CRMA*) (Case 1).

(1) Rank	(2) Case no.	(3) Case similarity score	(4) Prediction accuracy	(5) Predicted value filtered by CRMA	(6) Predicted value filtered by CRMA*
1	14	78.11154265	99.99375477	**48768.21026**	**48768.21026**
2	12	71.70764795	91.22976467	53048.61007	53048.61007
3	34	64.56916180	91.46362858	**44607.96057**	**44607.96057**
4	18	61.06650458	85.19326843	55992.68511	55992.68511
5	7	60.78816529	67.80337316	64473.95547	64473.95547
6	17	59.25953752	73.51594796	35854.65128	35854.65128
7	35	56.73599641	56.83200654	27717.68348	27717.68348
8	80	56.31537633	65.18956444	31793.76945	31793.76945
9	54	56.07303768	54.62480813	26641.20509	26641.20509
10	9	55.74396919	59.99333689	29259.50400	29259.50400
11	36	55.60761936	78.91750834	38489.06013	38489.06013
12	3	54.54470905	78.63022637	59193.56317	59193.56317
13	2	53.85588679	96.63033778	**47127.82954**	**47127.82954**
14	19	53.67200777	67.91878818	33124.84615	33124.84615
15	4	53.60675706	84.86738844	56151.62088	56151.62088
16	40	53.28100751	64.68848831	31549.38832	31549.38832
17	37	53.28024847	78.31581972	38195.60903	38195.60903
18	78	53.12230124	51.09288032	24918.63953	24918.63953
19	53	52.70142619	94.1708934	**45928.32763**	**45928.32763**
20	23	52.19718708	70.0397174	34159.24997	34159.24997
:	:	:	:	:	:

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) 36,734.62 ⩽ MRA ⩽ 49,892.14; (2) 42,120.60 ⩽ ANN ⩽ 51,758.34; (3) 42,120.60 ⩽ CRMA ⩽ 49,892.14; and (4) 40,042.65 ⩽ CRMA* ⩽ 52,353.49.

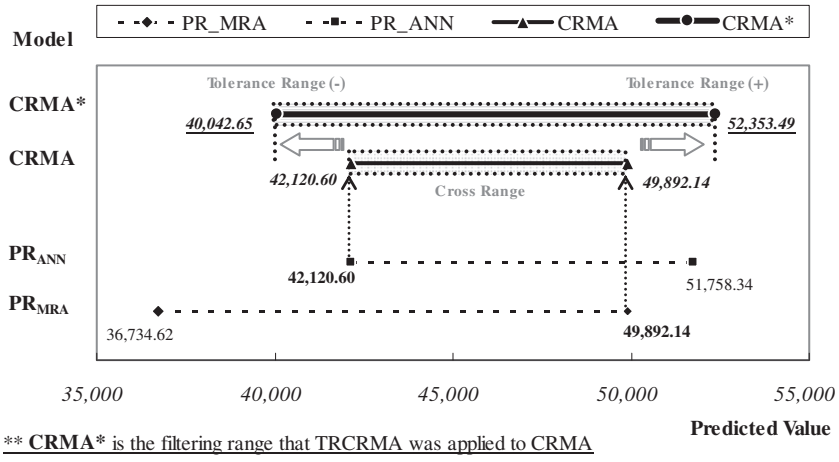

Fig. 2. Concept of CRMA* for filtering the predicted value.

1 using the MRA model (PV_{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 {= 43,313.38 × [1 − (15.18875240 ÷ 100)]} ⩽ PR_{MRA} ⩽ 49,892.14 {= 43,313.38 × [1 + (15.18875240 ÷ 100)]} 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:

$$PV_{ANN} \times (1 - (SER_{ANN}/100)) \leqslant PR_{ANN} \leqslant PV_{ANN}(1 + (SER_{ANN}/100)), \quad (4)$$

where PV_{ANN} is the predicted value of the ANN model, 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 (PV_{ANN}) was deduced as 46,939.47 with the software program Neuro Solutions 5. The standard error rate of the ANN model (SER_{ANN}) was calculated as 10.26612982%. Thus, the predicted range of the ANN model (PR_{ANN}) was deduced as 42,120.60 {= 46,939.47 × [1 − (10.26612982 ÷ 100)]} ⩽ PR_{ANN} ⩽ 51,758.34 {= 46,939.47 × [1 + (10.26612982 ÷ 100)]} 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):

$$Max(Min(PV_{MRA}), \quad Min(PV_{ANN})) \leqslant CRMA \leqslant Min(Max(PV_{MRA}), \quad Max(PV_{ANN})), \quad (5)$$

where PV_{MRA} is the predicted value of the MRA model, PV_{ANN} is the predicted value of the ANN model, and *CRMA* is the cross-range between the predicted value of MRA and the ANN model. For example, the predicted value of Case 1 in the MRA model (PR_{MRA}) was

deduced as $36{,}734.62 \leqslant PR_{MRA} \leqslant 49{,}892.14$ using Eq. (3). The predicted range of the ANN model (PR_{ANN}) was deduced as $42{,}120.60 \leqslant PR_{ANN} \leqslant 51{,}758.34$ using Eq. (4). Thus, *CRMA* was deduced as 42,120.60 [= Max (36,734.62, 42,120.60)] $\leqslant CRMA \leqslant$ 49,892.14 [= Min (49,892.14, 51,758.34)] using Eq. (5) (refer to Fig. 2).

2.2.2.4. 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:

$$Min(CRMA) \times (1 - TRCRMA/100) \leqslant CRMA \leqslant Max(CRMA) \times (1 + TRCRMA/100), \quad (6)$$

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 $42{,}120.60 \leqslant CRMA \leqslant 49{,}892.14$ 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 (= $42{,}120.60 \times (1 - (4.9333423 \div 100))$) $\leqslant CRMA^* \leqslant$ 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:

$$f_{SER}(x) = \frac{|V_{Test\_Case} - PV_{Retrieved\_Case}|}{V_{Test\_Case}} \times 100, \quad (7)$$

where f_{SER} is the function of the standard error rate, V_{Test_Case} is the value of the *Test_Case*, and $PV_{Retrieved\_Case}$ is the predicted value of the *Retrieved_Case*.

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

$$f_{PA}(x) = 100 - f_{SER}(x), \quad (8)$$

where f_{PA} is the function of the prediction accuracy and f_{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% [= abs $(48{,}771.26 - 48{,}768.21) \div 48{,}771.26 \times 100$]. When f_{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

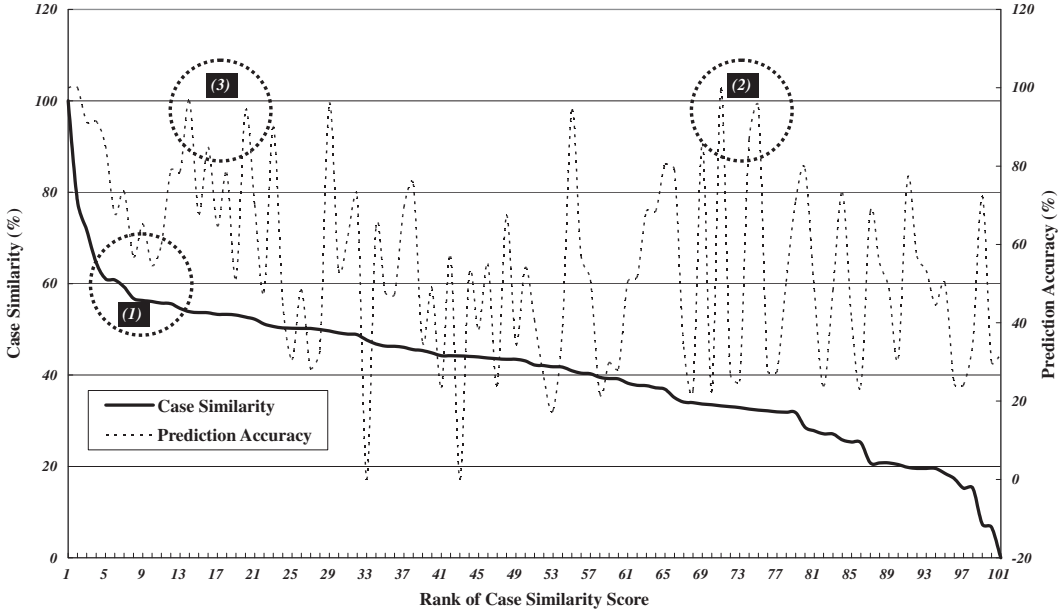

Fig. 3. Correlation between the case similarity score and the prediction accuracy (Case 1).

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. Thus, 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 $42{,}120.60 \leqslant CRMA \leqslant 49{,}892.14$, which were retrieved using Eq. (5) (refer to (5) predicted value filtered by CRMA in Table 2) and another four cases were filtered with $40{,}042.65 \leqslant CRMA^* \leqslant 52{,}353.49$, 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 $15{,}300.44 \leqslant CRMA \leqslant 16{,}386.93$, which was the result of Eq. (5), did not filter any case as shown in the fifth column [(5) Predicted Value filtered by CRMA], $14{,}545.61 \leqslant CRMA^* \leqslant 17{,}195.36$ 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 3
The similar cases produced by filtering engines (i.e., RCS and CRMA*) (Case 72).

(1) Rank	(2) Case no.	(3) Case similarity score	(4) Prediction accuracy	(5) Predicted value filtered by CRMA	(6) Predicted value filtered by CRMA*
1	73	96.17991949	70.08053452	11575.38623	11575.38623
2	75	94.71121284	73.94187771	12213.17444	12213.17444
3	71	93.70760918	81.08078006	13392.32572	13392.32572
4	67	89.29874029	83.42134734	13778.92338	13778.92338
5	69	82.48743226	86.87114090	18685.79124	18685.79124
6	64	79.48230739	88.21675882	14570.99411	**14570.99411**
7	63	77.85311445	91.24237993	15070.74390	**15070.74390**
8	70	74.72900797	69.10483470	11414.22733	11414.22733
9	84	67.59786276	60.34882027	23066.55270	23066.55270
10	46	67.56670899	78.84165989	20012.04174	20012.04174
11	22	67.08935344	81.64234757	13485.08131	13485.08131
12	76	65.89479349	55.53575086	23861.54004	23861.54004
13	68	60.68032115	59.05701028	23279.92436	23279.92436
14	47	60.37531078	57.33042681	23565.10869	23565.10869
15	77	60.24002012	51.00820591	24609.36655	24609.36655
16	81	57.75052784	96.48560101	17097.74557	**17097.74557**
17	83	55.98643206	59.02520781	23285.17726	23285.17726
18	56	55.79534094	71.04307138	11734.37098	11734.37098
19	44	55.21443021	70.78694335	11692.06564	11692.06564
20	20	54.42260658	83.42671891	13779.81061	13779.81061
:	:	:	:	:	:

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) $15{,}300.44 \leqslant MRA \leqslant 20{,}780.71$; (2) $13{,}335.58 \leqslant ANN \leqslant 16{,}386.93$; (3) $15{,}300.44 \leqslant CRMA \leqslant 16{,}386.93$; and (4) $14{,}545.61 \leqslant CRMA^* \leqslant 17{,}195.36$.

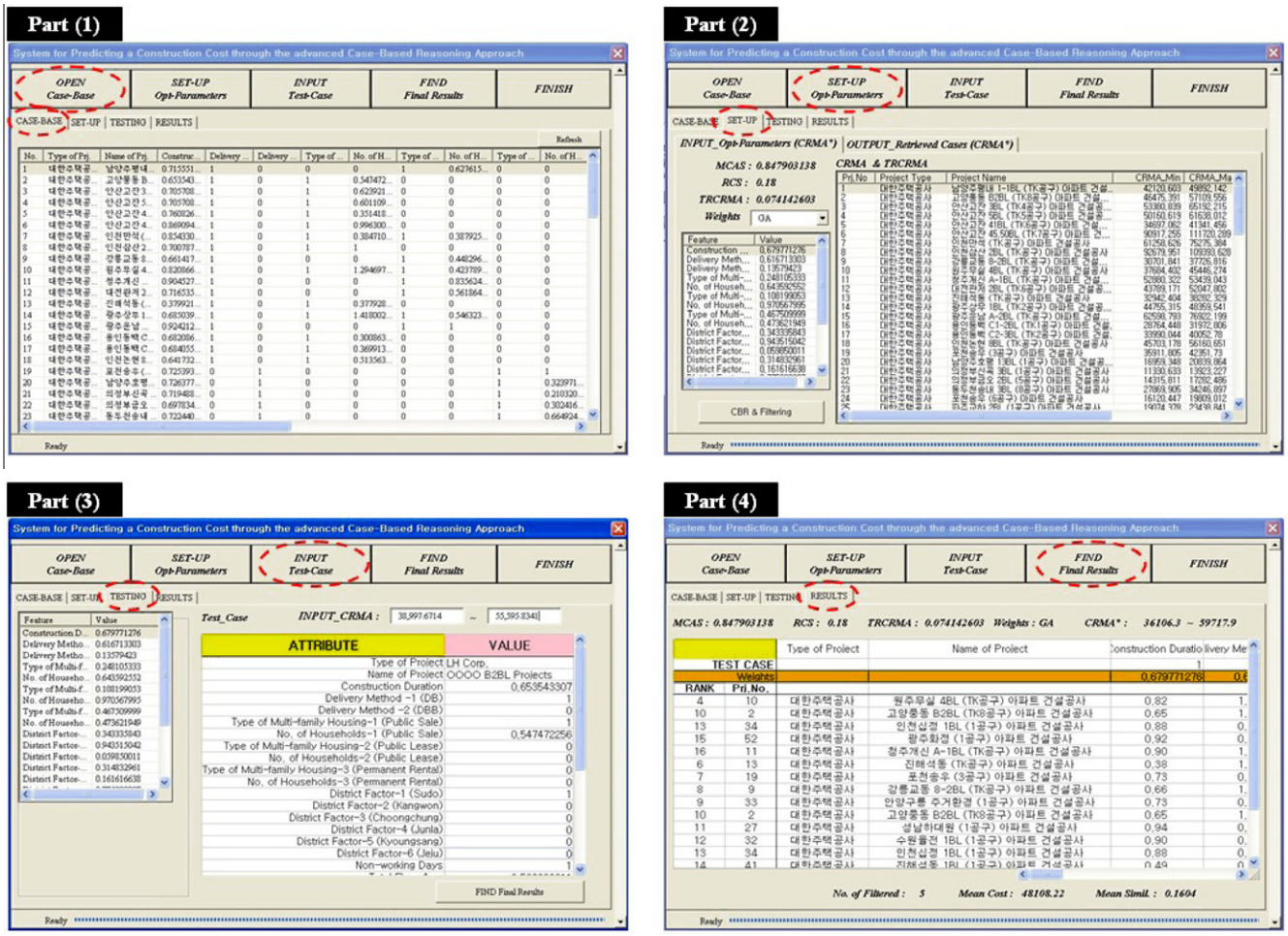

Fig. 4. Graphical user interface of the advanced CBR model.

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.

(1) Methodology	(2) Attribute weight	(3) No. of cases	(4) Mean	(5) Standard deviation	(6) Median	(7) Min.	(8) Max.	(9) 5th Percentile
CBR-1 (Koo et al., 2009) including MCAS and TAW as optimization parameters	Feature Counting	99	77.092	17.923	79.357	4.677	99.981	43.12
	MRA(orig.)	99	76.388	21.267	80.181	4.415	99.424	42.46
	MRA(abs)	100	78.235	19.405	82.906	4.415	99.265	48.10
	ANN	99	81.078	18.475	86.702	3.652	99.981	54.90
CBR-2 (Koo et al., 2010) including MCAS and RAW as optimization parameters	GA	100	83.540	15.920	88.479	4.678	99.699	54.27
CBR-3 (this study) Including MCAS, RAW, RCS, and TRCRMA as optimization parameters	GA	95	87.428	9.066	89.692	47.264	99.099	74.284

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

Table 5
Detailed information on the retrieved cases using the advanced CBR model (Case 1).

Variable	Case no.	Construction duration	Delivery method	Type of multi-family housing	No. of households	Site location	Non-working days	Total floor area	No. of stories above the ground		No. of stories below the ground		Size of household		Land ratio	Case similarity score	Total construction cost	Prediction accuracy
									Min.	Max.	Min.	Max.	Min.	Max.				
Cost	Test case	1	727	Design-build	Public lease	1050	District 1 (Sudo)	90	101,539.238	16	20	1	2	48	59	14.69		48,771.3	
	Retrieved case	14	696	Design-build	Public sale/public lease	23/914	District 4 (Junla)	75	105,358.319	16	20	-	0.5	49	84	18.43	78.11	48,768.2	99.99
		34	899	Design-bid-build	Public sale/public lease	696/224	District 1 (Sudo)	90	94,609.000	13	25	-	1	46	84	17.11	64.57	44,608.0	91.46
		2	664	Design-build	Public sale	888	District 1 (Sudo)	115	107,356.551	17	20	-	1	75	84	13.90	53.86	47,127.8	96.63
		53	713	Design-bid-build	Public lease	1148	District 4 (Junla)	75	143,297.000	15	20	-	0	84	84	16.24	52.70	45,928.3	94.17
																	Construction cost prediction results	46,608.1	95.57

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

Arditi, D., & Tokdemir, O. (1999). Comparison of case-based reasoning and artificial neural networks. *Journal of Computing in Civil Engineering, 13*(4), 162–169.

Attalla, M., & Hegazy, T. (2003). Predicting cost deviation in reconstruction projects: Artificial neural networks versus regression. *Journal of Construction Engineering and Management, 129*(4), 405–411.

Construction Industry Institute (CII), 1998. *Improving early estimates best practices guide*. Research report 131-11, Univ. of Texas, Austin, Tex.

Dogan, S. Z., Arditi, D., & Gunaydin, H. M. (2006). Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management, 132*(10), 1092–1098.

Duverlie, P., & Castelain, J. M. (1999). Cost estimation during design step: Parametric method versus case based reasoning method. *Advanced Manufacturing Technology, 15*(12), 895–906.

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management, 124*(3), 210–218.

Kim, G., Kim, S., & Kang, K. (2004). Comparing accuracy of prediction cost estimation using case-based reasoning and neural networks. *Journal of Architectural Institute of Korea, 20*(5), 93–102.

Koo, C., Hong, T., Hyun, C., & Koo, K. (2009). A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. Paper scheduled for publication in the Canadian Journal of Civil Engineering.

Koo, C., Hong, T., Hyun, C., Park, S., & Seo, J. (2010). A study on the development of a cost model based on the owner's decision making at the early stages of a construction project. *International Journal of Strategic Property Management, 14*(2), 121–137.

Korea Institute of Construction Technology (KICT), (2007). In Lee, Y. (Ed.), *Construction cost index handbook*. Gyeonggi, Korea.

Korea National Housing Corporation (KNHC), (2005). In S. Jang (Ed.), *Construction cost analysis handbook on multi-family housing*. Gyeonggi, Korea.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management, 132*(7), 750–758.

Mark, W., Simoudis, E., & Hinkle, D. (1996). Case-based reasoning: Expectations and results. In D. B. Leake (Ed.), *Case-based reasoning: Experiences. Lessons, and future directions*. Cambridge, MA: AAAI Press/MIT Press.

Michael, R. M. (1990). *Improving the accuracy of early cost estimates for federal construction projects*. Washington, DC: National Academy Press.

Phaobunjong, K. (2002). *Parametric cost estimating model for conceptual cost estimating of building construction projects*. PhD thesis, Univ. of Texas, Austin, Tex.

Rifat, S. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. *Canadian Journal of Civil Engineering, 31*(2), 677–683.

Watson, I. (1997). *Applying case-based reasoning: Techniques for enterprise systems*. San Francisco, California: Morgan Kaufmann Publishers, Inc.