Doctor of Philosophy

Dual-Optimization Method for Improving Accuracy in GA-CBR Cost Estimating Model

February 2017

Department of Architecture & Architectural Engineering
The Graduate School
Seoul National University

Sooyoung Kim

유전알고리즘-사례기반추론 공사비 예측 모델의 정확도 향상을 위한 듀얼 옵티마이제이션 방법

Dual-Optimization Method for Improving Accuracy in GA-CBR Cost Estimating Model

指導敎授 李 鉉 秀

이 論文을 金洙瑛의 博士學位論文으로 提出함

2016年 10月

서울大學校 大學院

建 築 學 科

金 洙 瑛

金洙瑛의 博士學位論文을 認准함

2016年 12月

委 員 長 井 紋 緖 (인)

副委員長 李 鉉 秀 (인)

委 員 지 석 호 (인)

委 員 손 보 식 (인)

委 員 유 정 호 (인)

Dual-Optimization Method for Improving Accuracy in GA-CBR Cost Estimating Model

A dissertation submitted to the Graduate School of
Seoul National University
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

by
Sooyoung Kim

December, 2016

Approval Signatures of Dissertation Committee

Moonseo Park

Moonseo Park

Hyun-Soo Lee

Hyun-Soo Lee

Seokho Chi

Bosik Son

Bosik Son

Jungho Yu

Abstract

Dual-Optimization Method for Improving Accuracy in GA-CBR Cost Estimating Model

Sooyoung Kim

Department of Architecture & Architectural Engineering
The Graduate School of Seoul National University

As large amount of time and resources used in completing a construction project, cost estimating is consistently carried out in the all of the stages. In particular, the early stage of cost estimating is very important in the decision-making as it determines the success and failure of the project. Case-based reasoning is widely used as an effective methodology for early cost estimation in construction projects. It has the advantages that it can infer persuasive and accurate answers relatively fast, easy to maintain, and the accuracy increases as it used. So many researchers have been conducting research to improve the accuracy and usability of cost estimating models by using case-based reasoning.

The accuracy of the case-based reasoning cost estimating model has been affected by retrieving and adaptation. Case retrieval has a significant effect on the performance of CBR models. Retrieval accuracy depends on how many attributes are used, what kinds of attributes are used, and how the attribute weights are assigned in a model. However, existing methods used only a limited number of qualitative variables by applying a subjective weight assignment method or excluded these from their model. This can limit the number of variables which can be used, or the difference between qualitative attributes can be disregarded. Additionally, since the problem description is not completely same as previous cases, old solutions should be adjusted to fit new situations. There are several adaptation methods for CBR, it is necessary to improve the estimation accuracy by applying suitable adaptation methods for the cost estimating model. The other method of adaptation is using several cases for problem solving. If only one case is used to solve the problem, it cannot reflect the good traits of other similar cases. The weighted mean method is most widely utilized because it can reflect the difference of case similarities to deduce a solution by giving higher weights to more similar cases. However, there is a disadvantage that the difference of weights is calculated relatively small.

In an effort to address these problems, this research attempts to present the dual-optimization method for considering qualitative variables in CBR

cost estimating models based on a genetic algorithm. This method can assign not only the attribute weights, but also the quantified values of the qualitative attributes. This method is able to apply more attributes than existing methods through quantification of the qualitative variables. Additionally, this research suggested two adaptation methods for the GA-CBR cost estimating model. The retrieving error adaptation is the method to calculate the differences and to adjust the solutions of retrieved cases. By reflecting estimation error caused by differences between target and retrieved cases, a retrieved solution is adjusted to be more appropriate. Furthermore, the improved weighted mean method was suggested to alteration method for multiple cases adaptation. By assigning weights according to the similarity distribution of retrieved cases, the improved weighted mean method can increase the influence of more similar cases.

To validate the proposed methods, three kinds of validations were conducted to estimate the construction cost of military barracks and public apartment projects. The results of validation 1 indicate that the dual-optimization method is improved in terms of accuracy and stability compared to previous methods. In validation 2 and 3, the estimation accuracy is increased when the proposed adaptation methods are used to the GA-CBR cost estimating model. These validation results support that the

proposed methods can be utilized for construction cost estimation to make better decision.

This research has significant in that it suggests three new methods to improve the disadvantages of existing methods. Consequently, the new GA-CBR cost estimating model can make more accurate cost estimation compared to other methods. It is expected that the proposed cost estimating model and methods will support stakeholders of construction project to make better decision at early stage of project. Moreover, the dual-optimization is a general-purpose method; it is expected to be more readily applied to a problem of other fields. In the dual-optimization, despite of its excellent performance, the more qualitative variables and values are utilized, the longer the length of the chromosome to be optimized and the longer the calculation duration. This research hopes that this problem will be solved by advance of computer science and improvement of its algorithm.

Keywords: Cost estimation, Case-based Reasoning, Case retrieving, Case adaptation, Optimization, Genetic Algorithm
Student Number: 2009-23032

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Research Backgrounds

Construction projects have a lot of uncertainty results from their diversity and uniqueness. Moreover, large amounts of time and resources are required to complete construction projects. In order to solve various problems arising from construction projects and to reduce the uncertainty of projects, cost estimations are consistently performed at all stages of projects. In particular, at an early stage, conceptual and preliminary cost estimation is very important for decision-making, as it determines the success or failure of the project (Trost and Oberlender 2003, Koo et al. 2011). Furthermore, early-stage cost estimations are especially important because cost changes increase as the project proceeds, thus decreasing its cost-effectiveness (Duverlie and Castelain 1999). However, very limited information about construction is available for conceptual and preliminary cost estimation during the early stages of a project because details are not yet settled (Hong et al. 2011, Kim and Kim 2010, Arafa and Alqedra 2011). This makes it difficult to estimate construction cost accurately during the early stage.

The traditional cost estimating methods include the cost index method, the cost capacity method, the unit price estimating method, and the parametric cost estimating method. Although these methods are relatively simple, their prediction accuracy is relatively low because they cannot reflect various factors affecting the construction cost. As the advance of computer science, many methods using artificial intelligence technology have been used to improve the estimation accuracy. Above all, since case-based reasoning models produce persuasive solutions by applying a similar reasoning process to that of human problem solving when limited information is provided; this method is widely used as an effective methodology for early cost estimation in construction projects (Arditi and Tokdemir 1999, Kim and Kang 2004, Ji et al. 2011). Compared with other cost estimation methods, case-based reasoning has the advantages that it can infer persuasive and accurate answers relatively fast, easy to maintain, and the accuracy increases as it used because self-learning is possible. So many researchers have been conducting research to improve the accuracy and usability of cost estimating models by using case-based reasoning.

1.2 Problem Statement

As mentioned before, early stage cost estimation is crucial for construction project success. The information that can be obtained at the early stage of a project is very limited. In this regard, case-based reasoning is highlighted as an effective method of early cost estimation, utilizing knowledge by comparing past experience and then generates solutions. Case-based reasoning is composed of four processes: retrieve, reuse, revise, and retain. In order to solve a problem, similar cases in the case base are retrieved by using retrieving methods. Then, solutions of retrieved cases are reused to solve the current problem by copying or adapting other similar cases. The proposed solution is applied to solve the problem. The proposed solution is revised according to the problem solving results. Finally, the solution is retained into the case base as a new case.

In order to improve the estimation accuracy, previous CBR researches have been focused on following two challenging issues. The first issue is how to retrieve the similar cases (Aamodt and Plaza 1994, Leake 1996, Luu et al. 2005). Retrieval is a fundamental investigation process for appropriate reasoning, and thus has a significant effect on the performance of CBR models (Goh and Chua 2009). Most of the CBR cost estimating models use

the nearest neighbor retrieval method for retrieving similar cases because it is suitable for numerical information, easy to develop, and less affected by missing or incorrect information than other retrieval methods. In nearest neighbor retrieval, retrieval accuracy depends on how many attributes are used, what kinds of attributes are used, and how the attribute weights are assigned in a model.

Previously, many researchers have tried to find the best way to calculate attribute weights for CBR. Kim and Kang (2004), Doğan et al. (2006), and Dikmen et al. (2007) used a gradient descent approach; Chun and Park (2006) and Ji et al. (2010) adopted regression analysis; Ji et al. (2011) applied genetic algorithm (GA); and so forth. Although these attempts developed mathematics-based weight value assignment methods, they did not consider qualitative properties or used non-quantitative scale factors in a limited way in their models. In this context, Doğan et al. (2006) gave Boolean parameters to the textual information that could not reflect the difference of each attribute weight. Park and Han (2002) and An et al. (2007) used an analytic hierarchy process whose results are prone to changes according to the expert group used. Ji et al. (2011) tried to quantify ratio scale variables by applying fuzzy functions; however, this approach is suitable only for certain attributes whose values can be distinguished by other factors, and needs another attribute which can be used to estimate.

The second issue is how to adapt the solution of retrieved cases (Goh and Chua 2009, Liao et al. 2012, Hu et al. 2015). Since the problem description is not completely same as previous cases, old solutions should be adjusted to fit new situations. There are several adaptation methods for CBR, it is necessary to improve the estimation accuracy by applying suitable adaptation methods for the cost estimating model.

In the CBR cost estimating model, the most commonly used way of adaptation is that adjust the solutions by reducing the effect on difference between a problem and retrieved cases. Garza and Maher (1999), Liao et al. (2012) used GA; Lotfy and Mohamed (2002), Policastro et al. (2003) adopted artificial neural network; Policastro et al. (2008), Sharifi et al. (2013) applied support vector machine; Ji et al. (2010) utilized regression analysis; Qi et al. (2012) used decision tree for case adaptation. Although these methods can obtain reliable adaptation results, these have the disadvantage of requiring additional analysis or new database or self-learning for the adaptation.

The other method of adaptation is using several cases for problem solving. If only one case is used to solve the problem, it cannot reflect the good traits of other similar cases (Begum et al. 2009, Hu et al. 2015), so many researches applied multiple case adaptation methods to generate a

new solution. There are several methods for multiple case adaptations such as equal mean, median, and weighted mean. Among them, the weighted mean method is most widely utilized because it can reflect the difference of case similarities to deduce a solution by giving higher weights to more similar cases (Kim and Kim 2010, Hu et al. 2015). Although the weighted mean method can adapt solutions of retrieved cases based on the degree of case similarity; however, there is a disadvantage that the difference of weights is calculated relatively small. This happens very frequently because retrieved cases generally have relatively high similarity values and the difference between weights has decreasing as the number of reusing cases increases.

1.3 Research Objective

As elaborated in chapter 1.2, although there are several methods for calculating attribute weights of CBR cost estimating model, they used only a limited number of qualitative variables by applying a subjective weight assignment method or excluded these from their model. This can limit the number of variables which can be used, or the difference between qualitative attributes can be disregarded. These problems can reduce the accuracy of the cost estimation.

In an effort to address these problems, this research attempts to present an alternative method for considering qualitative variables in CBR cost estimating models based on a genetic algorithm that can find the global optimum solution. The proposed new method, the dual-optimization method, configures the combination of attribute weights and random valuables, which correspond to the values of qualitative variables, as a chromosome, performing simultaneous optimization using a genetic algorithm. This method can assign not only the attribute weights, but also the quantified values of the qualitative attributes. This method is able to apply more attributes than existing methods through quantification of the

qualitative variables. Thus, it is possible to retrieve more similar cases, leading to more accurate cost estimates.

Additionally, to further improve the accuracy of cost estimating, the following two adaptation methods for reuse process are suggested and applied to the dual-optimization cost estimating model. Firstly, the retrieving error adaptation method is proposed. As above mentioned, the retrieved cases are not entirely consistent with the problem description, so there is a need to adjust the retrieved cases to suit the condition of the problem. This research defines the error caused by the differences between the problem and the similar cases as a retrieving error, and proposes the method to calculate the differences and to adjust the solutions of retrieved cases. This method is able to adapt the residuals of retrieved cases without additional analysis or database so the adaptation can be conducted quickly and accurately with little effort.

Next, the improved weighted mean method is proposed that utilized the value of standard normal cumulative distribution for the similarity of retrieved cases. This method makes it possible to express the importance of the retrieved cases by utilizing the position relative to the average of the similarity. By assigning weights according to the similarity distribution of

retrieved cases, the improved weighted mean method can increase the influence of more similar cases.

Finally, this research suggests a new GA-CBR cost estimating model with three improved methods above mentioned. By improving the limitations of existing models, the GA-CBR cost estimating model is able to retrieve more similar cases in retrieving phase, and is able to modify the retrieved solutions to be more suitable for the problem. Consequently, this makes it possible to further improve the accuracy of cost estimation.

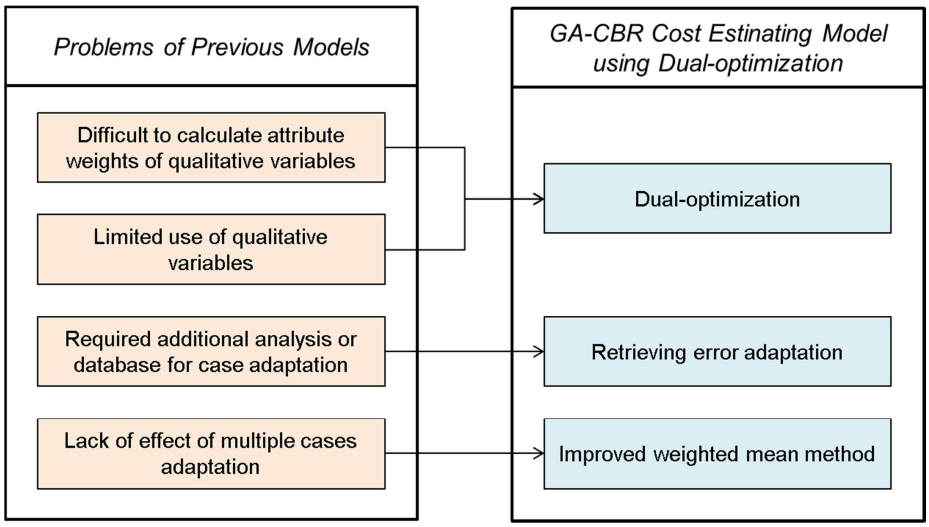


Figure 1-1 Existing Problems and New GA-CBR Cost Estimating Model

1.4 Research Scope and Process

To achieve the research goals, the research scope is limited to the early-stage cost estimation. Estimation during early phase of the project is very important because it is critical to the initial decision making. The more the project is progressed, the less the possibility of reducing the final cost. Among the CBR Process, this research is focused on retrieve and reuse phase because the accuracy of cost estimating model is highly dependent on the completeness of these phases. The other phases of the CBR; revise and retain, are correlation with the maintenance and learning of CBR cost estimating model. Accordingly, to validate the proposed methods, this research utilizes two kinds of cases; military barracks and public apartment projects conducted in Korea.

Figure 1-2 summarizes the research process of this research. This research has begun with addressing research backgrounds, problem statement, research objective, and research scope and process. Two challenging issues of CBR cost estimating model are clarified in problem statement, and developing a new CBR cost estimating model to improve estimation accuracy is settled to research objective. In chapter 2, it has reviewed literature on early stage cost estimation, case-based reasoning,

genetic algorithms, and types of variables. This research identifies the implications and limitations of existing CBR methods, focusing on retrieving and adaptation methods. In chapter 3, the logic of the dual-optimization method, which assigns the attribute weights by quantifying the values of qualitative attributes based on the genetic algorithm, is explained in detail. Additionally, two adaptation methods, the retrieving error adaptation method and the improved weighted mean method, are suggested to improve estimation accuracy in the GA-CBR cost estimating model. In chapter 4, the GA-CBR cost estimating models of two kinds of cases are developed based on the proposed methods, and cost estimating process is described in detail. In chapter 5, three kinds of validations are conducted to verify applicability and validity of proposed methods by comparing CBR outputs. Finally, conclusions are offered in chapter 6.

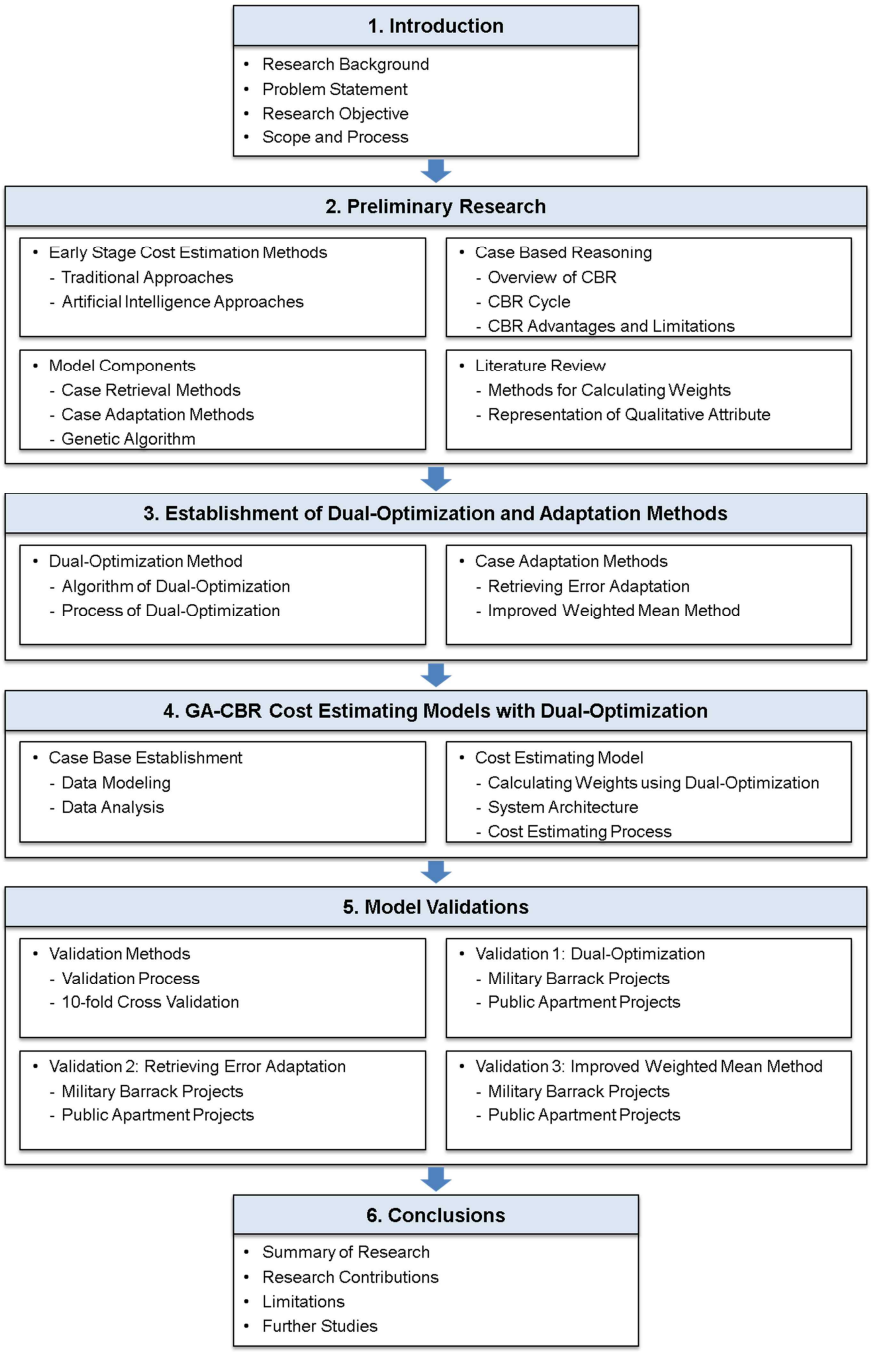


Figure 1-2 Research Process

Chapter 2. Preliminary Research

In chapter 2, this research explains academic backgrounds of GA-CBR cost estimating model. This research first introduces historical development of early stage cost estimation methods. Traditional and artificial intelligence approaches for early stage cost estimating are briefly explained, and why CBR can be helpful for cost estimating at early stage is discussed. Then, principles of CBR are elaborated to focus on its problem solving process and the advantages and limitations of CBR are discussed. Next, model components required for developing GA-CBR cost estimating model are presented. Case retrieval and adaptation methods in CBR, principles of genetic algorithm, types of variables, and construction cost index are reviewed in detail to find out of directions of developing the GA-CBR cost estimating model. Finally, literature reviews on calculating weights and representing qualitative attributes are conducted to describe the limitations of previous researches and the need for developing the new method.

2.1 Early Stage Cost Estimation Methods

A cost estimate is the approximation of the project. For construction projects, cost estimates should be made several times over the process because the completion of the construction project takes much time and resources. Moreover, the construction projects have a lot of uncertainty resulted from their diversity and uniqueness. Cost estimates can be classified into three methods by the phase; conceptual and preliminary estimate, detailed estimate, and definitive estimate Conceptual and preliminary cost estimates are established in the initial phase of the project. It is used to set a construction budget and these estimates are shown to the owner whether the project will be financially responsible or profitable economically in the initial phase. Once entering into preliminary and detailed engineering phase, the estimates and budgets are updated constantly to incorporate new information and match the management objective through a variety of estimation methods. This feeds back into the design phase in order to complete the entire project to fit the budget (Barrie and Paulson 2000).

Figure 2-1 illustrates the level of influence and information on project costs. Comparing with a whole expenditure of a project, the relative

expenditure of the initial phases of the project is very small. Similarly, project information available utilized to make a decision is also restrict in detail. Contrastively, decisions made during the early phases have influence on future expenditures on the project. In other words, estimates during early phase of the project is very important because it is critical to the initial decision making, the more the project is progressed the less the possibility of reducing the final cost.

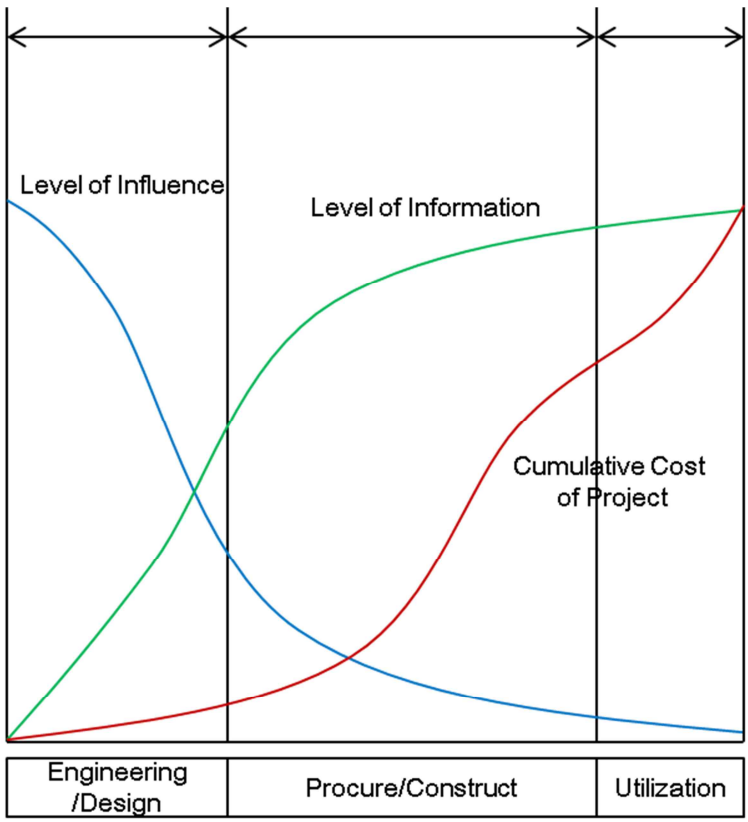


Figure 2-1 Level of Influence and Information (revised from Paulson 1976)

When compared to the other cost estimation method, the expected accuracy of the cost estimation at the initial phase is relatively low because the information of project is restricted. Table 2-1 shows the classification matrix for cost estimate made by AACE (Christensen and Dysert 2005). According to the table, expected accuracy range of class 5 corresponded to the conceptual and preliminary phase is from +30/-20% to +100/-50%, relatively. When level of project definition is increased, expected accuracy range is decreased, on the other hand, preparation effort for estimating will increase as more information is available. Conceptual and preliminary estimates are used to determine the feasibility of a project quickly or screen several alternative designs. Also, when the range of expected accuracy is larger, the more contingency should be needed. So a variety of approaches for conceptual and preliminary estimates have been developed and applied to increase the accuracy of estimating.

Table 2-1 Cost Estimate Classification Matrix from AACE (revised from Christensen and Dysert 2005)

Estimate Class	Lever of Project Definition	End Usage	Methodology	Expected Accuracy Range	Preparation Effort (relative)
Class 5	0% to 2%	Concept screening	Capacity factored, parametric models, judgment, or analogy	L: -20% to -50% H: +30% to +100%	1
Class 4	1% to 15%	Study or feasibility	Equipment factored or parametric models	L: -15% to -30% H: +20% to +50%	2 to 4
Class 3	10% to 40%	Budget, authorization or control	Semi-detailed unit costs with assembly level line items	L: -10% to -20% H: +10% to +30%	3 to 10
Class 2	30% to 75%	Control or bid/tender	Detailed unit cost with forced detailed take-off	L: -5% to -15% H: +5% to +20%	5 to 20
Class 1	65% to 100%	Check estimate or bid/tender	Detailed unit cost with detailed take-off	L: -3% to -10% H: +3% to +15%	10 to 100

2.1.1 Traditional Approaches

Most of the existing early stage cost estimating methods can be classified as follows: cost index method, cost capacity method, unit price estimating method, and parametric cost estimating method. They are listed in ascending order of accuracy, and in ascending order of cost and complexity. Cost index method is the method to estimate a reproduction cost by using the index for the building cost released from reliable institutions such as Engineering News-Record (ENR). The method can reflect changes in time, place, technology, and productivity (Barrie and Paulson 2000). Cost capacity method can be applied to changes in size of projects of similar types (Ahuja et al. 1994). It reflects the nonlinear increase in cost with size, and sometimes used together with the cost index methods. Unit price estimating method is a single price method based on the cost per functional unit of the building such as a bedroom of hotel. The unit price is multiplied by the required quantity and then all costs are summed to calculate the total cost of building. This method is applicable to the buildings those of functional unit are to be repeated, such as hotels and hospitals. One of the most accurate methods for estimating the construction cost is parametric estimating. In order to predict future costs, the method uses the historical cost data and statistical techniques. The implicit

assumption of parametric cost estimating is that the same forces that affected the past cost will affect the future cost. The major advantage of a parametric estimating is that the estimate can be conducted quickly and is easily replicated. Parametric models calculate the dependent variables of cost and duration based on one or more independent variables that reflect the characteristics of projects. These methods, although the advantage of being simple to estimate, the prediction accuracy is relatively low because it does not reflect various factors affecting the construction cost.

2.1.2 Artificial Intelligence Approaches

With ever changing computer scientific advances, Artificial Intelligence (AI) techniques have been developed and adopted by various engineering disciplines. Many approaches using AI techniques have been attempted for a more effective estimating the construction cost at the preliminary phase. Rule-based reasoning (RBR) is one of the successful methods in the field of AI. RBR extracts all the rules of the problem areas from human experts, and then the rules are organized implement the rule base, and works out solutions by reasoning. However, there are some problems in RBR. Actually, it is often not possible to establish all the rules beforehand while solving the problem. When the problem does not match the rules, it is difficult to solve the problem. Thus, in order to solve the problem, the knowledge (rules) must be supplemented again and again. Furthermore, the performance of RBR decreases with increasing the number of rules because the rules related to solve the given problem applies in sequence.

To solve the problems of RBR, several alternative AI techniques have been developed and utilized to cost estimating. One of the commonly used AI techniques for cost estimating is Artificial Neural Networks (ANN). ANN is the statistical learning algorithms inspired by biological neural

networks. Hegazy and Ayed (1998) developed an artificial neural network model for parametric cost estimation of highway projects. Elhag and Boussabaine (1998) developed neural network models to predict the tender price of school buildings. Likewise, Günaydın and Doğan (2004) attempted to apply ANN for estimating cost of structural system of buildings in Turkey. Despite the excellent performance, however, cost estimating using ANNs still has many disadvantages. The major problem of ANN is that cannot identify a causal relationship between input and output variables. In other words, the knowledge acquisition process is a black box. ANN is more acceptable in a problem that only few knowledge is present, than cost estimating. In addition, there is no mechanism for variable selection in ANN, and it may result in overfitting when increasing the number of variables. Also, it takes a long time due to high computational complexity.

In order to overcome the disadvantages, case-based reasoning (CBR), which is another AI technique, has been attracted much attention as an alternative method for cost estimating. CBR is the process of solving problems by recognizing their similarity between a problem and past cases that were used to solve the problems (Riesbeck and Schank 1989). CBR is more flexible compared to ANN in updating the system and it is more successful in handling missing information. In this overall perspective, CBR is more successful to deal with construction related problems, which

is experience oriented (Arditi and Tokdemir 1999). In recent years, numerous studies have attempted to apply the CBR method for cost estimating. Yau and Yang (1998) developed the method for estimating construction duration and costs of building construction project at the preliminary design stage. Doğan et al. (2006) proposed the CBR model for early cost prediction of structural system. An et al. (2007) developed the CBR model for estimating the construction cost of residential building using Analytic Hierarchy Process (AHP).

The other AI technique, genetic algorithm (GA), is utilized for assistant methods to other methods rather than used independently. Generally, GA is often used to solve the global optimization problem. Kim and Kang (2004) developed the ANN model incorporating a GA in estimating construction costs. Kim and Kim (2010), Hong et al. (2011), Ji et al. (2011), and Koo et al. (2011) used GA to the optimization process of the CBR model. Additionally, various early stage cost estimating methods have been developed by using other AI techniques such as fuzzy logic.

2.2 Case Based Reasoning (CBR)

2.2.1 Overview of CBR

Case-based Reasoning (CBR) is a branch of AI appeared to recognize that humans look at things on the basis of past experience. Aamodt and Plaza (1994) defined CBR as the act of remembering similar past situations to resolve a new problem and reusing information and knowledge. Watson and Marir (1994) defined CBR as the act of applying similar past successful best practices to a problem to solve it. To apply CBR, it is important to determine whether a past experience can be applied to the present problem. CBR remembers the experience or knowledge learned from past problems as cases. When faced with a new problem, CBR models select a most similar case to the problem from the library; apply the solution of the case for the problem, and come up with an answer. The primary difference between CBR and other artificial intelligence methods is that CBR uses detailed information from previous cases for problem solving rather than making rules. Secondly, when a problem is solved, it is stored as a past case and can be used in the future when a similar problem arises (Kolodner 1993, Aamodt and Plaza 1994).

Leake (1996) lists four assumptions that represent the basis of the CBR as follows. First, the same actions performed under the same conditions will tend to result in similar (Regularity). Second, works or experiences tend to be repeated (Typicality). Third, small changes under any circumstances are only a small change in the solution or in the interpretation (Consistency). Finally, the difference among tends to be reduced when the work is repeated, and small difference is easy to balance (Ease of Adaptation). Figure 2-2 represents how the assumptions are used to solve problems in CBR. Given a new problem, the problem is described in terms of previous problems, and the most similar previous problems can be found. To propose a solution, some adaptation is required because the new problem does not fully correspond to the most similar problem. Based on the distance between the new problem and the previous problem, a new solution can be created by adaption. This additional problem-solution case adds in the CBR model and it will be used to solve other new problems in the future.

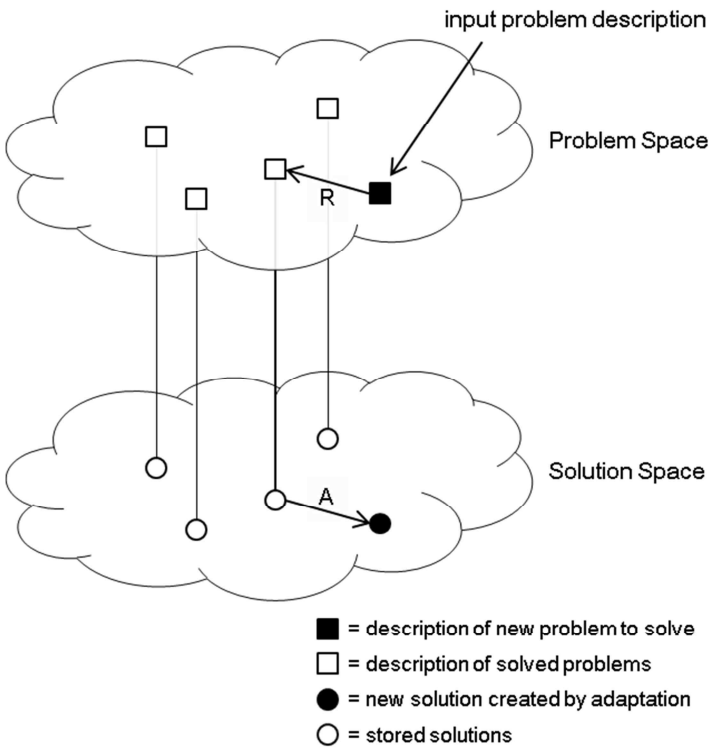


Figure 2-2 Relationship between Problem and Solution Space in CBR (adapted from Leake 1996)

CBR has been found to be appropriate for areas hard to find certain rules for problem solving; in particular, if it can lead to effective decision-making from past experience, it is a very effective methodology for solving the problem (Turban 1992). Currently, CBR is used widely in various construction fields where decision making is necessary, particularly in architecture design, method selection, scheduling, cost estimation, and other areas where past knowledge and experience are useful.

2.2.2 CBR Cycle

A general CBR cycle is divided into the following four processes (Aamodt and Plaza 1994, Watson 1999):

1) **Retrieve** the most similar case comparing the case to the library;

2) **Reuse** the retrieved case to solve the current problem;

3) **Revise** the proposed solution if necessary;

4) **Retain** the solution as a part of new case.

Figure 2-3 shows the basic model of the problem solving cycle in CBR. When performing the reasoning, a new problem is matched against cases in the library and one or more similar cases are retrieved. Next, a solution suggested by the matched cases is reused and the success of the solution is tested. If the retrieved cases do not match the target situation closely, the solution will have to be revised. Finally, the solution is retained as new cases in the library.

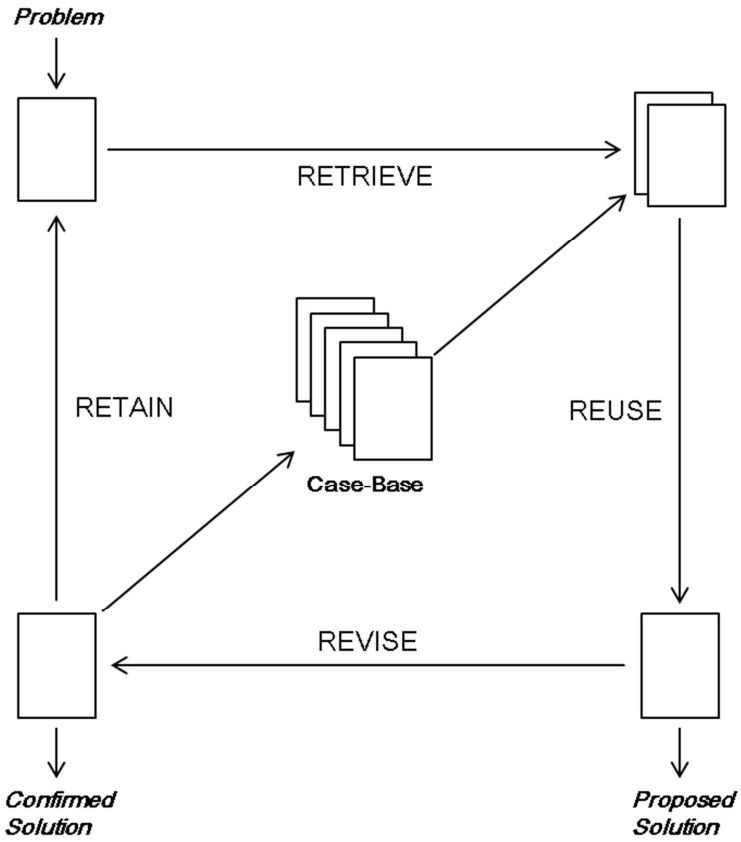


Figure 2-3 CBR Cycle (adapted from Watson and Marir 1994)

Case retrieval is a process that retrieves the most similar cases to a new problem. The retrieval process is critical to the success of CBR (Montazemi and Gupta 1997, De Mantras et al. 2006). To retrieve the most similar cases, first, a set of relevant problem descriptors is identified. This involves to filter out noisy problem descriptors, to infer other relevant problem features, to check whether the feature values make sense within the context, to

generate expectations of other features. Then, cases in the library are matched to the target problem and a set of sufficiently similar cases are returned. Finally, the best case is selected from the set of similar cases.

Case reuse is a process that proposes the solution for a new problem from the solutions in the retrieved cases. Reusing the retrieved case solution in the context of the new case focuses on: identifying the differences between the retrieved and the current case; and identifying the part of a retrieved case which can be transferred to the new case. Case reuse is performed in two ways: Copy and Adapt. When the difference between the new case and the retrieved cases is small or negligible, the solution of the retrieved case is transferred to the new case directly as its solution case (Copy). However, when using the adaptation, the differences are taken into account that the solution from case base is adapted in accordance with the new situation of a given problem.

The solution generated by reuse is applied and evaluated to the real situation. If the solution is successful in solving a given problem, it is stored in the case base. When the solution proves incorrect, it is necessary that revises the case solution generated by the reuse process. The reason for the failure should be explained and the solution is repaired by detecting and adjusting the errors. This provides an opportunity to learn from failure.

Case retain is the process of incorporating whatever is useful from the new case into the case library. This involves deciding what information to retain and in what form to retain it; how to index the case for future retrieval and integrating the new case into the case library. Whether the generated case proves successful or not during the revise phase, related information should be retained as the useful information for new reasoning in the future.

2.2.3 CBR Advantages and Limitations

CBR retrieves the most similar problem with the current problem in the memory, adapts the previous solutions to suit the current problem by considering the difference between problems. Knowledge is used by the CBR as past cases differently from RBR, this technique is to simply reuse past cases similar to the given problems to solve current problems. According to Shiu and Pal (2004), advantages of using CBR can be summarized as follow:

- Reduce the knowledge acquisition task
- Avoid repeating mistakes happened in the past
- Provide flexibility in knowledge modeling
- Reason in domains that have not been fully understood, defined, or modeled.
- Forecast the possible success of a suggested solutions
- Acquire continuous knowledge acquisition over time
- Reason in a domain with a small body of knowledge
- Possible to reason with inadequate or inaccurate data and concepts
- Reason with incomplete or imprecise data and concept
- Avoid repeating all the steps that need to be taken to arrive at a solution

- Provide a means of explanation
- Can be used in many different ways
- Can be applied to a broad range of domains
- Reflect human reasoning

One of the major advantages of CBR is that is not only simpler by utilizing the knowledge of past cases than to answer with conventional reasoning methods, but also possible to reason in an area with the problem that have not been well-structured through similar cases. In other words, CBR has applicability to address a new problem by using the past problems with similar features. CBR also has flexibility to search because the partial combination is allowed. The second advantage of CBR is that, if the past cases exist, allows saving time and cost for obtaining the solution. When solving a problem, there is often unable to build all the knowledge in advance. However, building new knowledge need a lot of time and cost. Whereas, CBR can derive the solutions without any special reasoning if the given problem is similar to experience gained in the past. Using CBR can save the effort for solving problem, especially in an area with the complex problem. The other advantage of CBR is that does not limit the ability to learn. CBR can acquire knowledge from past experience automatically, does not cost much time and money since the acquisition of knowledge as well as the increasing knowledge are carried out automatically. These

advantages have a greater significance since it overcomes the disadvantages of the other AI techniques such as RBR and ANN. Table 2-2 compares the features of RBR and CBR.

Despite its great advantages, CBR still involves some limitations and challenges. One of the major disadvantages of CBR is their high memory requirements. As the number of cases increases, the problem solving time also increases due the processing necessary to answer the queries. These problems may lead to increase system costs and reduce system performance. Yet, these problems become less important due to advance of computer science. Another limitation concerns the problem domains. Basically, CBR solves problems relying on past practices. However, CBR systems have difficulties in handing new problems that did not existing in the past. Although CBR is very useful for repeated problems, the new input from the expert is required to solve a completely different problem.

Table 2-2 Comparison between RBR and CBR

	Rule-Based Reasoning	Case-Based Reasoning
Unit of knowledge	· Rules	· Cases
Knowledge acquisition	· Extract rules from experts	· Collect cases
Construction of the system	· Difficult to convert knowledge to rules	· Establish the case base by indexing and collecting cases
Maintenance	· Require a lot of efforts in the expansion and maintenance of the system	· Easy to maintenance
Learning	· Repeat same mistakes	· Learning from both successes and failures
Explanation of results	· Difficult to justify the solution	· Explanation becomes easier and pervasive

2.3 Model Components

2.3.1 Case Retrieval Methods

Case retrieval is a process that a retrieval algorithm retrieves the most similar cases to the current problem. Case retrieval requires a combination of search and matching. For this, two retrieval techniques are used by the major CBR applications: nearest neighbor retrieval algorithm and inductive retrieval algorithm.

1) Nearest Neighbor Retrieval

Nearest neighbor algorithms are the most commonly used techniques in CBR (Watson 1999, Patterson et al. 2003). It is a method that involves finding the most similar case among existing cases using a similarity index. To retrieve the most similar case, definitions of attributes that can represent the characteristics of cases and a criterion with which to compare the values of the attributes are necessary. Figure 2-4 shows a flowchart for nearest-neighbor retrieval. In order to extract similar cases using nearest-neighbor retrieval, case similarity must first be defined. Establishing case similarity is defined as the similarity between an estimation object and previous cases.

It can be considered as the act of establishing an appropriate similarity function, which is an attempt to handle the hidden relationships between objects (Burkhard 2001). Although the evaluation of case similarity can vary depending on the researcher, when applying nearest-neighbor retrieval, the general method is to calculate the sum of the products of the attribute similarities and their weights. Hence, the evaluation of similarity depends on the function that calculates the similarity and its attribute weight. A typical evaluation function is used to calculate nearest neighbor matching as shown in the following equation (Watson 1999):

$$Similarity \quad (T,S) = \sum_{i=1}^{n} f(T_i, S_i) \times w_i$$

(Eq. 2-1)

where T is the target case, S is the source case, n is the number of attributes in each case, i is an individual attribute from 1 to n, f is a similarity function for attribute i in case T and S, and w is the importance weighting of attribute i.

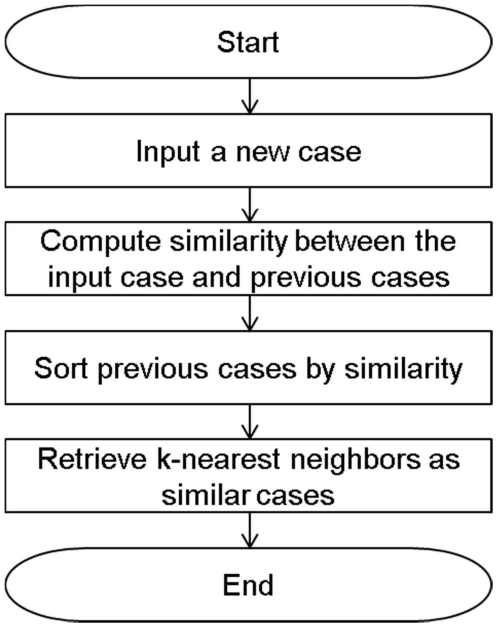


Figure 2-4 Process of Nearest Neighbor Retrieval

Figure 2-5 represents a simple scheme for nearest neighbor retrieval. In this 2-dimensional space, k is user-defined constant, and when a new problem is given, the retrieval algorithm searches the nearest k cases to the problem in the space. If k is too small, the good trait of other similar cases can be ignored. In contrast, if k is too large, the estimation accuracy can be decreased because even dissimilar cases are utilized for problem solving. There are some techniques for selecting a good k, such as a hyper-parameter optimization; however, this research limits the value of k to five because the scope of this research is focused on calculating attribute weights and quantifying qualitative variables.

Nearest neighbor retrieval depends on the distance function. For every point x, y, and z, the distance function D should satisfy the following properties:

1) non-negativity: $\mathrm{D}(\mathrm{x},\mathrm{y}) \geq 0$.

2) reflexivity: $\mathrm{D}(\mathrm{x},\mathrm{y}) = 0$ if and only if $x = y$.

3) symmetry: $\mathrm{D}(\mathrm{x},\mathrm{y}) = \mathrm{D}(\mathrm{y},\mathrm{x})$.

4) triangle inequality: $D(x,y) + D(y,z) \geq D(x,z)$.

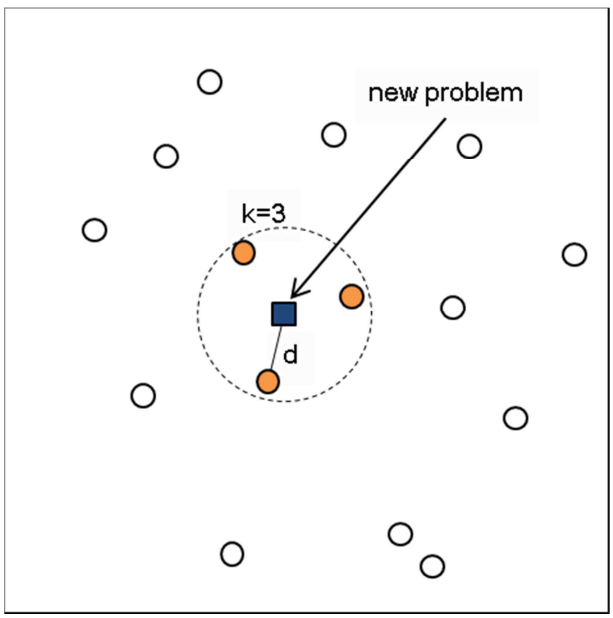


Figure 2-5 Simple Scheme for Nearest Neighbor Retrieval

A commonly used distance function for continuous variables is Euclidean distance, as shown in the following equation:

$$D_u(x,y) = \sqrt{\sum_{i=1}^{n}(x_j - y_j)^2}$$

(Eq. 2-2)

where D_u is the Euclidean distance between x and y, n is the number of attributes in each case, and j is an individual attribute from 1 to n.

A variant of Euclidean distance, called the weighted Euclidean distance, is used when having some idea of the relative importance of each variable. The weighted Euclidean distance is represented as the following equation:

$$D_w(x,y) = \sqrt{\sum_{i=1}^{n}\frac{w_j(x_j - y_j)^2}{\sum w_j}}$$

(Eq. 2-3)

where D_w is the weighted Euclidean distance between x and y, n is the number of attributes in each case, j is an individual attribute from 1 to n, and w_i is the weight of attribute i.

Generally, the similarity and the distance are represented by a value between 0 and 1. When two points are exactly equal, and sum of weights is

1, the distance between two points is 0 and the similarity is 1. When using the weighted Euclidean distance, the similarity is represented as the following equation:

$$Similarity \quad (x, y) = 1 - D_w(x, y) = 1 - \sqrt{\sum_{i=1}^{n} \frac{w_j (x_j - y_j)^2}{\sum w_j}}$$

(Eq. 2-4)

Prior to using the above equation, each variable values should be normalized. This prevents to give a small weighting for the attributes with a small range, as opposed to give a large weighting for the attributes with a large range.

2) Inductive Retrieval

Inductive retrieval algorithm is a technique that determines which features do the best job in discriminating cases and generates a decision tree type structure to organize the cases in memory (Watson, 1999). A decision tree will retrieve the similar case with the decisions made in the input level searching the database. It is a hierarchical tree where the decision will be made once there is no sub tree is available. Inductive retrieval is best if the retrieval goal is well-defined. The cases are indexed according to major

influences and are retrieved using extraction rules or a decision tree (Barletta 1991, Brown and Gupta 1994). This approach is also useful when a single case feature is required as a solution, and when that case feature is dependent upon others. Figure 2-6 shows a typical decision tree for inductive retrieval. When a target case is given, the algorithm would traverse the decision tree and search for the best matching case in the case base. If A of the target case is true, the algorithm first selects the left branch. After this, the algorithm traverses to the node and selects next branch according that B1 is true or false. When B1 of the target case is true, Case 1 is retrieved as a similar case. For the target case, CBR retrieves and reuses the past solution of Case 1.

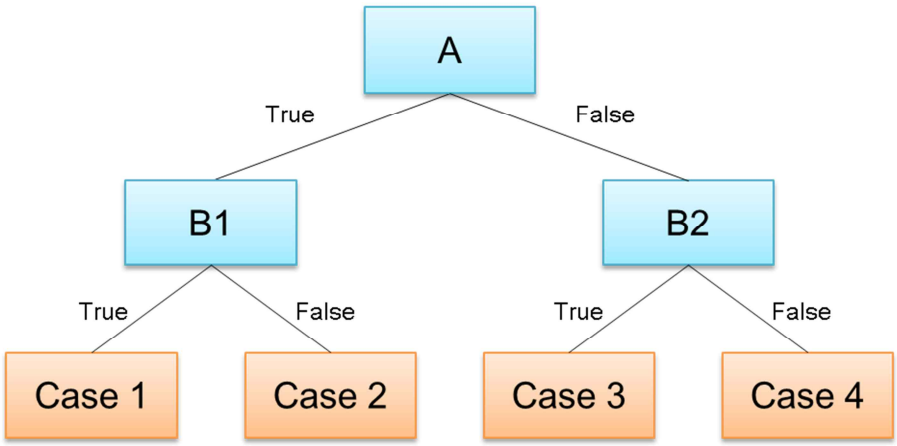


Figure 2-6 Decision Tree for Inductive Retrieval

3) Nearest Neighbor Retrieval vs. Inductive Retrieval

One of the advantages of nearest neighbor algorithm is that the entire case base is searched. Thus, making sure no vase has been overlooked. Moreover, nearest neighbor algorithm is easier to build than inductive algorithm. However, retrieval speed of nearest neighbor algorithm is relatively slow when searching over a large case library. Although inductive retrieval is quicker and requires less computational power because every single case in the case base does not need to be evaluated, it depends on pre-indexed case library, which is time-consuming process. Moreover, it is impossible to retrieve a case while case data is missing or unknown.

Construction projects are non-repetitive, and identical cases from the past are rare. Additionally, most of the variables used in construction cost estimation are expressed as numerical figures. Using such data as a foundation, nearest-neighbor retrieval is more effective than inductive retrieval in CBR construction cost estimating model. Furthermore, as it is affected less by missing or incorrect information than inductive retrieval, it can enhance the reliability of a cost estimating model. Hence, in this study, nearest neighbor retrieval is used for case retrieval in CBR.

2.3.2 Case Adaptation Methods

Since a new problem is not exactly the same as previous cases, old solutions should be adjusted to fit new situations. To reuse old solutions in CBR, it is required to change previous solutions, called adaptation. In general, there are three kinds of adaptation in CBR according to the amount of changes. Null adaptions apply retrieved solutions to current problem without adapting. Transformation-based adaptations are performed by using rules, and they are divided into two main categories; substitutional and structural adaptations. Substitutional adaptations replace some part of the retrieved solution and structural adaptations change the structure of the solution and reorganize the solution. Derivational adaptations replay the method of deriving the retrieved solution on the new problem (Richter and Weber 2013). According to Watson and Marir (1994), there are several adaptation techniques as following; they are listed in ascending order of complexity:

- Null adaptation
- Parameter adjustment
- Abtraction and respecialisation
- Critic-based adaptation
- Reinstantiation

- Derivational replay
- Model-guided repair
- Case-based substitution

These adaptation techniques can be used alone or in combination. These techniques are usually implemented to automatic adaptation by using the intelligent machine learning methods such as GA (Garza and Maher 1999, Liao et al. 2012), ANN (Lotfy and Mohamed 2002, Policastro et al. 2003), support vector machine (Policastro et al 2008, Sharifi et al. 2013), regression analysis (Ji et al. 2010) and decision tree (Qi and Peng 2012). The details vary, but the methods adjust the solutions by reducing the effect on difference between a problem and retrieved cases. Although these methods can obtain reliable adaptation results, these have the disadvantage of requiring additional analysis or new database or self-learning for adaptation.

In the reuse process, the use of only one retrieved case may ignore the impacts of other similar cases (Begum et al. 2009, Hu et al. 2015). So many researches applied multiple case adaptation methods for generate a new solution, which are classified into the equal mean, the median, the weighted mean, and the regression analysis. Among them, the weighted mean is the

most widely utilized method, which calculates the weighted average of the solution values of retrieved cases (Hu et al. 2015).

$$w_{kj} = \frac{S_k}{\sum S_k}$$

(Eq. 2-6)

where w_{kj} is the weight of j^{th} solution value of retrieved case k, and S_k is the similarity of retrieved case k.

2.3.3 Genetic Algorithm (GA)

1) Overview of GA

To calculate the attribute weights in a CBR the cost estimating model, it is necessary to find the attributes that affect the cost and then determine the weights of the attributes in relation to the cost. In this research, a genetic algorithm (GA) is used to calculate attribute weights in the CBR cost estimating model. This is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics (Gen and Cheng 2000). Based on techniques inspired by evolutionary biology, such as selection, crossover, mutation, and replacement, the optimum value of each attribute weight that describes the best relationship between the attribute and the construction cost is extracted. In other words, GA selects only the genes adapted to a given environment. They are crossed by each other, and are mutated randomly to transfer a superior genotype to the next generation. Therefore, it will remain genes more appropriate for a given environment while the evolution has continued.

Unlike traditional optimization algorithms, GA does not require the mathematical operation such as the derivative of the objective function. The

algorithm can be performed when the set of genes are represented by the fitness that represents a suitable level for the solution in the given problem. Compared to other optimization algorithms, GA has different characteristics that GA performs a parallel search while traditional algorithms start the search at one point. Over the evolution, genes of the each generation exchange information accumulated by previous generations and start searching for a new area. The direction of the search is determined probabilistically for each generation. For these reasons, GA can be a global optimization, and is less likely to fall in local optimal point. Although GA does not ensure that must find a global optimal solution of the problem, generally it can find a good enough solution in a short period of time.

In order to solve a problem, it is modeled by setting the chromosomes and the fitness function to evaluate the solution. The chromosome is set of parameters which define a proposed solution to the problem. The solutions to be optimized are expressed as a string of numbers, which is represented by a chromosome. The set of chromosomes is the population. In other words, the population is a group of elements that could be solutions. In GA, the population is selected to include a chromosome with a dominant trait, which is an approximate optimal value. Population size says how many chromosomes are in one generation. If there are too few chromosomes, GA has a few possibilities to perform crossover and only a small part of search

space is explored. On the other hand, if there are too many chromosomes, GA slows down. The fitness function is established to evaluate chromosomes in order to select dominant ones, which continues into the next generation. This can be compared as how living things are well adapted to the environment of the real world in accordance of their genetic characterization. In each generation, the fitness of every individual in the population is evaluated; the fitness is usually the value of the objective function in the optimization problem being solved.

2) Genetic Operators

The GA is composed of several key operators such as selection, mutation, crossover, and replacement. Selection is a process for selection chromosomes having a dominant trait from the parent population based on the results of the evaluation. Selection of the solution has a significant impact on the performance of the GA. According to which selection methods are used, the speed to reach a global optimal point may be slow, or the solution may converge to local optimal points. Selection is mainly performed in a probabilistic methods, such as roulette wheel selection (Holland 1975), ranking based selection (Baker 1985), and tournament selection (Goldberg 1989). Generally, the roulette wheel selection is commonly used that give a higher probability of being selected to the

chromosome in order of their fitness. In the roulette wheel selection, a selection probability pi is defined as the following equation:

$$p_i = \frac{f(s_i)}{\sum_{j=1}^{n} f(s_j)}$$

(Eq. 2-5)

when $f(s_i)>0$ and $i=1,2,\cdots,n$. where s is a solution (chromosome) and f(s) is a fitness of a solution s.

The chromosome having large fitness is more likely to be selected and to participate in the crossbreeding of the next step. Conversely, even as bad solutions have an opportunity to spread their good traits. Even if the best chromosome in the current generation may be the closest solution to a local optimal point, as opposed to the bad chromosome may be closer to a global optimal point.

Crossover is one of the key operators to create new chromosomes in the population. In a process for the recombinant chromosome of the real life, a part of the parent chromosomes are exchanged with each other by a particular location. With modeling this phenomenon, crossover operator of GA produces offspring chromosomes by intersecting the genes of the parents. Crossover operation can be variously performed such as single or

multi-point crossover, uniform crossover, and multi-parent crossover. If crossover is performed too much, the best chromosome in the current generation may be not preserved by losing their dominant genes through the crossing of the other chromosome. To prevent such a problem, the method for performing a probabilistic crossover is generally used in commercial GA software. When the setting of crossover rate is relatively low, it is more likely to converge to the local optimal point because the generation of offspring chromosome is reduced. Conversely, when setting high, even though less chance of falling into local optimal point, the search speed can be reduced. Unfortunately, no satisfactory theory exists to determine the appropriate level of crossover rate. Figure 2-7 shows an example of crossover in GA.

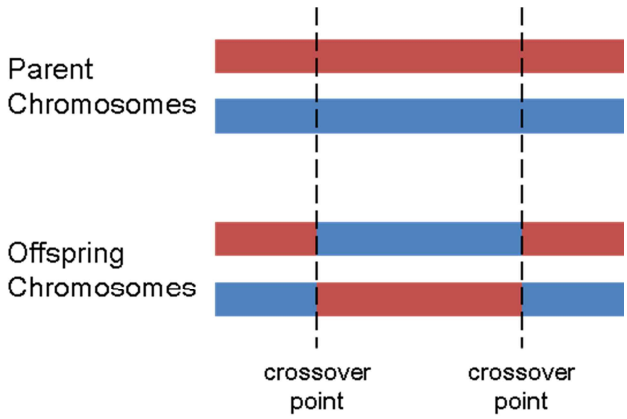


Figure 2-7 Example of Crossover in GA

The other key operator in GA is mutation. In ecological system, as well as crossing the genes, the probability to mutate genes to survive in a given environment also exists. Likewise, mutation operator changes the value or order of the genes in the chromosome arbitrarily. This is a type of localized random search for generating new chromosomes closed to existing ones. It is a method for maintaining diversity. To give a certain degree of randomness of the search process will be able to find a solution that cannot be accessed only by crossover. In GA, if the probability of mutation is too high, it may destruct the significant genes found during the search process, so the algorithm will be changed in a random search. Conversely, when mutation rate is set to low, it is hard to search other space than the initial population. Therefore, mutations in the proper rates can increase global searching performance and diversity of chromosomes. Figure 2-8 shows an example of mutation in GA.

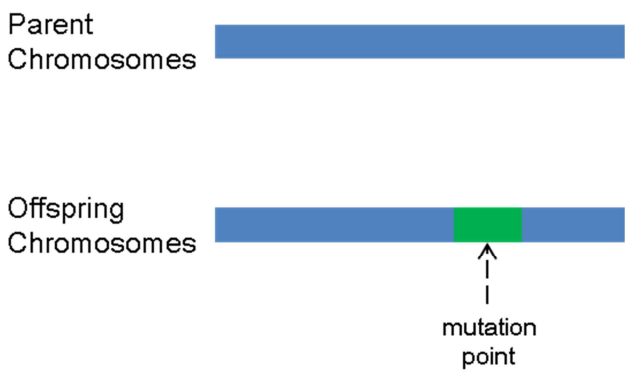


Figure 2-8 Example of Mutation in GA

Replacement is an operator that adds new chromosomes created through crossover and mutation into the population and removes inferior chromosomes in the existing population. Ways of replacement in GA are replacing the most inferior chromosome or replacing the most similar chromosome to the new chromosome.

3) Basic Process of GA

Figure 2-9 shows a flowchart for the basic process of GA. First, the initial population, which is a group of chromosomes, is generated randomly and all chromosomes are evaluated by the fitness function. In selection, two chromosomes that have higher fitness are selected from the population as parents to breed a new generation. In crossover, parts of genes are swapped between two parent chromosomes. In mutation, the values of genes are altered randomly. Through crossover and mutation, offspring chromosomes are created and are evaluated by the fitness function. When the offspring are better than the chromosomes in the population, the algorithm replaces the weaker members. This process will repeat until it reaches to a termination criterion (certain number of populations or improvement of the best solution). After termination of the process, the best chromosome in the population is extracted as the solution.

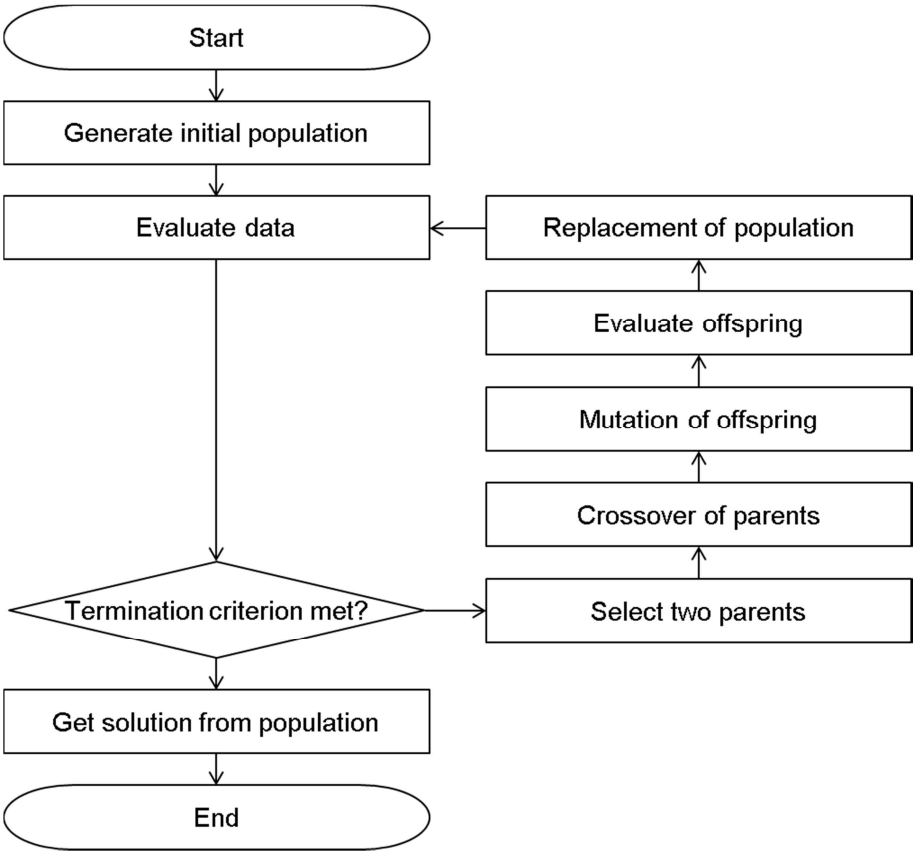


Figure 2-9 Process of Genetic Algorithm

2.3.4 Type of Variables

The variables used in statistics are divided into two types according to the scales of measurement, as shown in Table 2-3. A qualitative variable, also called a categorical variable, is a variable that is not numerical and is measured on a nominal or ordinal scale. A quantitative variable is a variable expressed in numerical values and is measured on an interval or ratio scale. Generally, in order to use mathematical techniques such as nearest-neighbor retrieval, variables should have an interval or ratio scale because nominal or ordinal scales are not able to express the interval between their values. Thus, before calculating the similarity between cases, it is necessary to quantify the qualitative variables, and it is important to select the right quantification method.

Table 2-3 Type of Variables according to Measurement Scale

Type	Scale	Definition	Example
qualitative variable	Nominal scale	a list of categories to which objects can be classified	· Roof type · Structure type
	Ordinal scale	a measurement scale that assigns values to objects based on their ranking	· Rank of preference · Finishing grade
quantitative variable	Interval scale	a measurement scale in which a certain distance along the scale means the same thing, but where "0" on the scale does not represent the absence of the thing being measured	· Likert scale · Celsius temperature
	Ratio scale	a measurement scale in which a certain distance along the scale means the same thing, and where "0" on the scale represents the absence of the thing being measured.	· Gross floor area · Kelvin temperature

2.3.5 Construction Cost Index

Construction cost will vary according to the time of the construction works. These variations of the construction cost not only make it difficult to predict the cost to be added in performing a new construction, but also make it difficult to identify the scale of construction economy. In this respect, the need for construction cost index has emerged. The construction cost index is an indicator of the construction cost price fluctuations based on the cost variation of the various components that make up the construction cost. By comparing the construction cost of a point in time with the cost of the reference point, the costs are represented with the same criteria.

In Korea, the Korean Construction Cost Index has published monthly by the Korea Institute of Civil Engineering and Building Technology (KICT) since 2004. Based on the weighted average prices of key resources spent on construction, the index has been published periodically by a specific point in time as 100. In order to calculate the index, the inter-industry relations table and the producer price index which are announced by the bank of Korea and the marker wages in construction which is announced by the Construction Association of Korea are utilized. The index can be effectively

useful to analyze cost fluctuations and market trends, to predict cost changing trends, and to adjust the cost data to the current price.

The construction cost index is utilized in two ways: normalization of cost data and converting cost to the current value. The data used for this research is from 2005 to 2008. In order to compare the data on the same criteria, it is necessary to normalize on a particular point in time. For these reasons, the data are normalized to 2008 by using the construction cost index. Table 2-4 shows the average annual Korean Construction Cost Index for the building construction: housing and the building construction: non-housing sectors.

Construction projects are being carried out for a very long time, and also have long period between the planning stage and the construction stage. Therefore, the estimated cost during the planning stage should be converted to a cost at construction time. In this respect, this research utilizes the construction cost index for converting the solution to current value as following equation.

$$C_{(current\ )} = C_{(year\ of\ solution\ \ )} \times \frac{CI_{(current\ )}}{CI_{(year\ of\ solution\ \ )}}$$

(Eq. 2-7)

where $C_{(current)}$ is the converted cost to current value, $C_{(year\ of\ solution)}$ is the original cost of solution, $CI_{(current)}$ is the cost index in current, and $CI_{(year\ of\ solution)}$ is the cost index in year of solution.

Table 2-4 Annual Average Korean Construction Cost Index

Year	305_Housing	306_Non-Housing
2000	60.97	58.19
2001	61.91	59.01
2002	65.14	61.85
2003	70.66	67.02
2004	75.48	72.40
2005	76.84	74.15
2006	78.43	76.06
2007	81.02	78.76
2008	92.71	92.19
2009	95.27	94.63
2010	98.67	98.77
2011	105.20	106.31
2012	109.31	109.41
2013	111.73	111.31
2014	114.53	113.23
2015	114.59	113.30

2.4 Literature Review

2.4.1 Calculating Weights in CBR Model

This research analyzed and summarized the aforementioned literature for CBR applications according to the objective, weighting method, number of quantitative and qualitative attributes, representation of qualitative attributes, and scale of qualitative attributes, as shown in Table 2-5. Current calculation methods used to determine the attribute weights in CBR mainly use gradient descent, regression analysis, analytic hierarchy processes, and GA. Although these methods can obtain reliable values based on a mathematical model, they have the following limitations.

The gradient descent approach is an optimization method that uses the gradient of a function. Yau and Yang (1998) used the gradient descent method for the construction time and cost estimating model. In the initial stage of apartment house cost estimation, Kim and Kang (2004) used the gradient descent method to calculate the attribute weight. However, the resulting value can vary according to the initial point, as it is prone to falling back into a local optimum.

Compared to other methods, regression analysis can determine relative weights in a short time. Chun and Park (2006) used regression analysis to calculate the weight from the financial quotient estimation model. Ji et al. (2010) used the standardization coefficient extracted from the regression analysis to calculate the attribute weight for public apartment cost estimation in its early stage. In regression analysis, it is assumed that the attributes are linearly independent; however, in reality, they are non-linear with a multi-collinearity relationship between the independent variables, leading to problems in the credibility of this method.

The analytic hierarchy process is a useful method when attributes are difficult to quantify or compare. Park and Han (2002) applied the analytic hierarchy process in the calculation of attribute weight for bankruptcy estimation model. An et al. (2007) used the analytic hierarchy process to calculate the attribute weights. Nevertheless, the weights can vary depending on the expert consulted, as the method is based on personal and subjective elements such as professional knowledge and experience. In contrast with the gradient descent method.

GA is a global search algorithm. Hence, they are less likely to fall back into a local optimum. Doğan et al. (2006) compares the performance of feature counting, gradient descent, and GA. As the result, the performance

when using GA is better than the performance when using the other two methods. Other researchers were also conducted researches to apply GA to CBR cost estimating models (Kim and Kim 2010, Hong et al. 2011, Ji et al. 2011), and the results have shown that the GA is suitable for calculating weights in CBR cost estimating models. Nonetheless, one weakness of this method is that calculations require a considerable time.

Table 2-5 Analysis of Previous CBR Models

Researcher (year)	Objective	Weighting method	Number of quantitative attributes	Number of qualitative attributes	Representation of qualitative attribute	Scale of qualitative attribute
Yau and Yang (1998)	Cost and duration estimation for building projects	Manual	9	1	Not described	Nominal
Park and Han (2002)	Bankruptcy prediction	Analytic Hierarchy Process	13	15	Binary	Ordinal
Kim and Kang (2004)	Cost of apartment housing prediction	Gradient Descent	6	4	Binary	Nominal
Chun and Park (2006)	Stock market index forecast	Regression Analysis	5	-	Not considered	-
Doğan et al. (2006)	Cost of structural system prediction	Feature counting, Gradient Descent, Genetic Algorithm	4	4	Binary	Nominal
An et al. (2007)	Cost of apartment housing prediction	Analytic Hierarchy Process	4	5	Binary	Ordinal
Dikmen et al. (2007)	Bid mark-up estimation	Gradient Descent	29	15	Binary	Nominal or Ordinal

Ryu et al. (2007)	Construction project planning	Not described	6	6	Binary	Nominal
Ji et al. (2010)	Cost estimation for apartment projects	Regression Analysis, Feature Counting	6	-	Not considered	-
Kim and Kim (2010)	Cost estimation for bridge projects	Genetic Algorithm	7	3	Binary	Nominal
Hong et al. (2011)	Cost estimation for multi-housing projects	Genetic Algorithm	13	6	Binary	Nominal
Ji et al. (2011)	Cost estimation for building projects	Genetic Algorithm	9	3	Binary or Converted by fuzzy method	Nominal or Interval

2.4.2 Representation of Qualitative Attributes

As summarized in Table 2-5, there have been many studies that have attempted to use a similar method to calculate attribute weights of qualitative attributes in CBR. One of the most frequently used approaches is to express similarity in binary parameters depending on whether the attribute values are exactly the same or not (Yau and Yang 1998, Park and Han 2002, Kim and Kang 2004, Doğan et al. 2006, An et al. 2007, Ryu et al. 2007, Kim and Kim 2010, Hong et al. 2011). In other words, if the attribute values appear identical, then the attribute similarity is 1; otherwise, it is 0, as the following equation (Doğan et al. 2006):

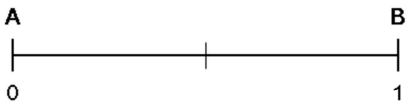
$$If\ \ T_j = X_{ij},$$
$$then\ S_{ij} = 1, or\ else\ \ S_{ij} = 0$$

(Eq. 2-8)

where T_j is the value of the attribute j of the target problem, X_{ij} is the value of the attribute j of the case i, S_{ij} is the attribute similarity of the attribute j of the case i.

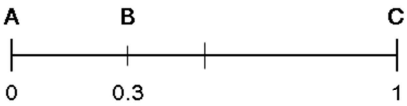
When applying nearest-neighbor retrieval, in principle the distance must be calculated using the interval or ratio scale. However, this approach is not

able to reflect differences between the attribute values of qualitative variables, as they utilize nominal or ordinal scales. If a qualitative variable consists only of two values such as presence or absence, the binary method is not a significant problem in the weight calculation. However, if there are three or more values in the qualitative variable, no matter the distance between the attribute values, the similarity is expressed only 0 and 1; it is not able to reflect the difference between the attribute values. As shown in the Figure 2-10, when there are two values, A and B, and if the value of the problem and the case are exactly same, the distance is 0 and the attribute similarity is 1, or else the distance is 1 and the similarity is 0. When there are three or more attribute values, if the attribute value of the problem is A, the attribute similarity of B should be greater than those of C because B is closer to A than C. However, although the distances of attribute values of A-B, A-C and B-C are 0.3, 0.7, and 0.4, respectively, the attribute similarities are just represented 0 regardless of the distance. That is, the binary method cannot reflect the difference between the three or more attribute values, which can reduce the accuracy of the cost estimation.



Attribute values		Distance	Attribute Similarity
A	A	0	1
A	B	1	0

(a) two values



Attribute values		Distance	Attribute Similarity
A	A	0	1
A	B	0.3	0
A	C	0.7	0
B	C	0.4	0

(b) three or more values

Figure 2-10 Attribute Distance and Similarity in Binary Method

In order to overcome these limitations, Ji et al. (2011) attempted to convert a qualitative variable into one with a ratio scale by using fuzzy methods. This approach can obtain reasonable attribute values with alteration functions. For the structure type attribute, they develop the structure type index as following equation:

$$Structure \quad type \quad index \quad (X_{SI}) = \frac{Quantity \quad of \quad vertical \quad formwork \quad (m^2)}{Unit \quad floor \quad area \quad (m^2)}$$

(Eq. 2-8)

Utilizing this index, the structure type attribute is converted to the ratio scale with the [0, 1] range as the following equation:

$$\mu_{RCcolumn}(X_{SI}) = f(x) = \begin{cases} 1, & X_{SI} < 0 \\ 2.451 - \dfrac{X_{SI}}{1.95}, & 2.83 \leq X_{SI} \leq 4.78 \\ 0, & 4.78 < X_{SI} \end{cases}$$

(Eq. 2-9)

However, the usage of this method is very limited, and is suitable only for certain attributes whose values can be distinguished by other factors. For example, although they converted structure-type attributes by utilizing the quantity of vertical formwork, it is difficult to apply this method to the military-type or roof-type attributes, as the alteration function is difficult to establish. Also, this approach needs another attribute which can be used to estimate, such as the quantity of vertical formwork.

2.5 Summary

Construction cost estimation is continuously carried out to fix the budget of project. As the project progresses, the amount of cost and information that can be obtained increases, but the impact of decision on the overall construction cost is reduced. Therefore, the cost estimation at early stage it is very important for project decision making. Unlike the other steps, the accuracy expected from the early stage cost estimation is relatively low. This means that the early stage cost estimation is very difficult due to the restricted information. If the accuracy of the project cost estimation is wider, it means that there are more uncertainties, which requires more contingency. In order to solve this problem, many studies have been conducted to improve the accuracy of the early stage cost estimation.

Traditional methods for estimating the construction cost at early stage include cost index method, cost capacity method, unit price estimating method, and parametric cost estimating method. Although these methods are relatively simple, their estimation accuracy is relatively low because they fail to take into consideration various variables affecting the construction cost. Hence, the methods using artificial intelligence techniques have been attempted to increase the cost estimation accuracy.

Rule-based reasoning is the basic artificial intelligence technique, however, there are disadvantages that it cannot solve the problem without the corresponding rules; it is very difficult to generate such a rule, and the reasoning speed decreases as the number of rules increases. Since artificial neural networks can solve problems with relatively high accuracy, many researches have been conducted to estimate the construction cost. However, there is a disadvantage that it is a black box model that cannot explain the relationship between input and output.

On the other hand, case-based reasoning is a method of retrieving the most similar cases through similarity between problems and past cases and using the solutions. Since case-based reasoning is easier to update than artificial neural network, and works well in situations even if information is lacking, it is an appropriate method for problem solving in construction field depended on experience. Case-based reasoning is consisted of four processes: Retrieve, Reuse, Revise, and Retain. In order to solve a problem, problem description is matched to previous cases in the case base; one or several similar cases are retrieved. The retrieved cases are reused for problem solving through adaptation, and if the result fails to solve the problem, the solution will be revised. Finally, problem solving result is retained to the case base. As such, the number of cases stored in the case

base is increased as the problem solving is repeated, and the accuracy of the reasoning is also increased.

The model components required for developing GA-CBR cost estimating model were reviewed. In case-based reasoning, nearest neighbor retrieval algorithm and inductive retrieval are mainly used for case retrieval. Among these methods, nearest neighbor retrieval, which is a method of retrieving similar cases based on the similarity between the problem and cases, is mainly used for cost estimation because it is effective for handling numerical data and is less affected by wrong or missing data. The nearest neighbor retrieval method depends on the distance function. In this study, the weighted Euclidean distance, which is an effective method when the importance of attributes is needed to consider, is used to calculate the similarity.

Since the past case is not exactly the same as the current one, the solution of the past case should be adapted to fit the new problem. There are many adaptation methods to mitigate the effects of differences between problems and cases, but they have the disadvantage of requiring additional analysis or database. In addition, as the use of only one case for problem solving may ignore the good characteristics of other similar cases, methods of deriving a solution through several cases have been used together.

Among them, the weighted mean is the most widely utilized method, which calculated the weighted average of the solution values of retrieved cases.

Finally, literature reviews on calculating weights and representing qualitative attributes were conducted. Various methods for calculating the weights of case based reasoning have been proposed. However, previous methods have the following limitations. Gradient descent method is likely to fall into the local optimum in the optimization process; regression analysis is affected by the nonlinearity and multi-collinearity of the construction cost variables; Analytic hierarchy process relies on subjective factors such as personal knowledge and experience, and has problems such as rank reversal. Genetic algorithm is less likely to fall into the local optimum due to the global optimization method, but it takes longer to compute than other methods. Looking at the issues related to the representation of qualitative attributes, the binary method, which expresses the attribute consistency as 0 and 1 without quantification, and the analytic hierarchy process are mainly used. However, these two methods have a disadvantage that it is difficult to reflect the difference of the values of qualitative attributes. As the fuzzy methods can obtain reasonable attribute values with alteration functions, however, it is suitable only for certain attributes whose values can be distinguished by other factors.

Chapter 3. Establishment of Dual-Optimization and Adaptation Methods

As above mentioned, there are two challenging issues for improving the estimation accuracy in case-based reasoning. In an effort to address the retrieving issue, this research proposes a new weights calculation method, dual-optimization. By assigning random variables to each value of qualitative attribute and optimizing these with the attribute weights, this method can assign not only the attribute weights but also the quantified values of the qualitative attributes. Additionally, to further improve the accuracy of cost estimating, the following two adaptation methods for reuse process are suggested and applied. The retrieving error adaptation method is suggested that defines the error caused by the differences between the problem and the similar cases as a retrieving error, and proposes the method to calculate the differences and to adjust the solutions of retrieved cases. Also, the improved weighted mean method is proposed that utilized the value of standard normal cumulative distribution for the similarity of retrieved cases.

3.1 Dual-Optimization Method

3.1.1 Algorithms of Dual-Optimization

In order to deal with the above-mentioned challenges regarding weight assignment for qualitative variables and quantifying variables, a new genetic-algorithm-based optimization method, named the dual-optimization method is suggested. Basically the dual-optimization method is developed on the foundation of an assignment method for CBR cost estimating models by Ji et al. (2011). The method uses the assumption that the cost of a specific case can be formulated by the sum of the products of attribute values of its weights. Under this assumption, the cost of each case can be represented as follows:

$$C_i = X_{i1}W_1 + X_{i2}W_2 + \cdots + X_{ij}\,W_j$$

(Eq. 3-1)

where C_i is the cost of the i^{th} case, X_{ij} is the value of the j^{th} attribute of the i^{th} case, and W_j is the weight of the j^{th} attribute.

For all cases in the case base, the set of linear equations can be converted into matrix form:

$$\begin{pmatrix} X_{11} & \cdots & X_{1j} \\ \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} \end{pmatrix} \times \begin{pmatrix} W_1 \\ \vdots \\ W_j \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_i \end{pmatrix}$$

(Eq. 3-2)

Before assigning the attribute weights by optimization, it is necessary to normalize the data because they are measured in different units and scales. In order to convert the data onto a scale of 0 to 1, it is assumed that all attribute values and costs are distributed normally and that the data are converted into values of the standard normal cumulative distribution according to the following equation:

$$p = F(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt$$

(Eq. 3-3)

where $X \sim N(\mu, \sigma^2)$, p is the normal cumulative distribution, X is a variable (cost or attribute), μ is the mean of a variable, and σ is the standard deviation.

After conversion, all parameters have the same scale in the range of 0 to 1. The above matrix (Eq. 3-2) is represented through the normalization as the following equation:

$$\begin{pmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_j \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_i \end{pmatrix}$$

(Eq. 3-4)

where c_i is the normalized cost of the i^{th} case, x_{ij} is the normalized value of the j^{th} attribute, and w_j is the weight of the j^{th} attribute.

Although this normalization method helps to reduce the effect of different variable ranges and units, it is not available for qualitative variables because these nominal and ordinal scale variables cannot be expressed by intervals between values. In this aspect, the previous research tried to convert qualitative variables onto a numerical scale in two ways: attributes are converted by the alteration function using a structure type index, and the other attributes of roof and hallway types are converted onto a binary scale where two values are rescaled to 0 or 1. As mentioned in Chapter 2, their trials still have the limitations that the binary method cannot represent the features of qualitative attributes in the way an interval scale can, and deriving a sort of fuzzy function for qualitative attributes is very difficult and need other attributes. Hence, the alteration functions can only apply for certain factors.

As an effort to improve these limitations, this research tries to expand the optimization target of the genetic algorithm by including the values of

qualitative attributes. In this context, Eq. 3-4 is redefined as the following equation when j-attributes are composed of m-quantitative attributes and l-qualitative attributes:

$$\begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{im} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} + \begin{pmatrix} q_{11} & \cdots & q_{1l} \\ \vdots & \ddots & \vdots \\ q_{i1} & \cdots & q_{il} \end{pmatrix} \times \begin{pmatrix} {w'}_1 \\ \vdots \\ {w'}_l \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_i \end{pmatrix}$$

(Eq. 3-5)

where c_i is the normalized cost of the i^{th} case, p_{im} is the normalized value of the m^{th} quantitative attribute of the i^{th} case, w_m is the weight of the m^{th} quantitative attribute, q_{il} is the value of the l^{th} qualitative attribute of the i^{th} case, and w'_l is the weight of the l^{th} qualitative attribute.

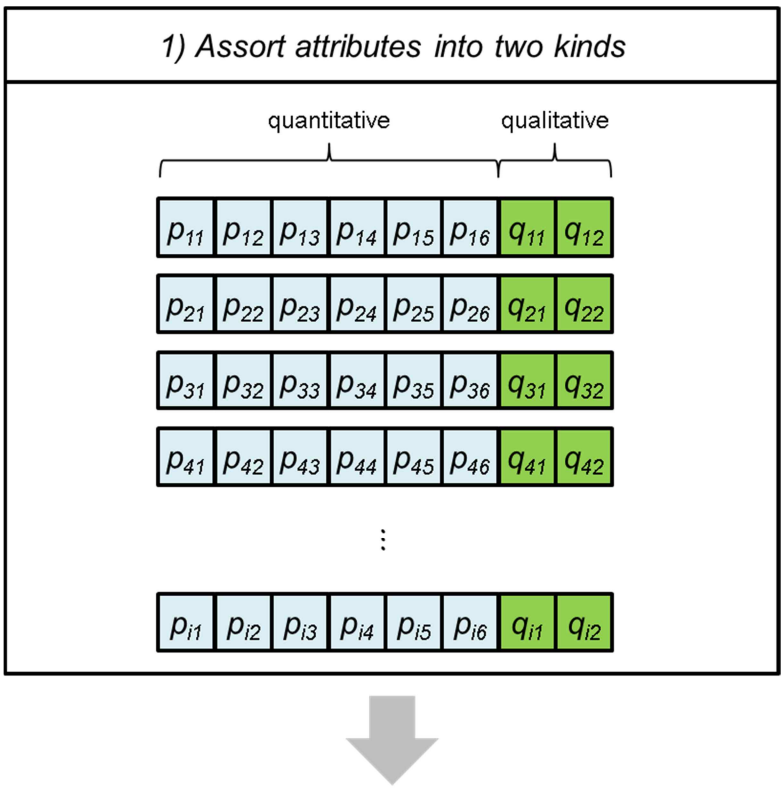
In order to quantify the value of qualitative attributes, a random variable corresponding to each attribute value of the qualitative attributes is utilized on the range of 0 to 1. As shown schematically in Figure 3-3, if the qualitative attribute 'roof style' consists of the three qualitative values a11, a12, and a13, and these are correspondingly defined as 'flat roof', 'gable roof', and 'sloped roof', these are converted into random variables b11, b12, and b13, respectively.

$when\ \ Q_l = \{x|x = q_l\} = \{a_{1l}, a_{2l}, \cdots, a_{nl}\},\ \ f(a_{nl}) = b_{nl},\ \ S(q_l) = b_{nl}\ \ and\ \ q_l = a_{nl},$

$r_l = S(q_l)$

(Eq. 3-6)

where, Q_l is the set of elements with the same qualitative value of the l^{th} attribute, a_{nl} is the value of the n^{th} same value of the l^{th} qualitative attribute, b_{nl} is a random variable corresponding to a_{nl},, and r_{il} is the quantified value corresponding to q_{il}.



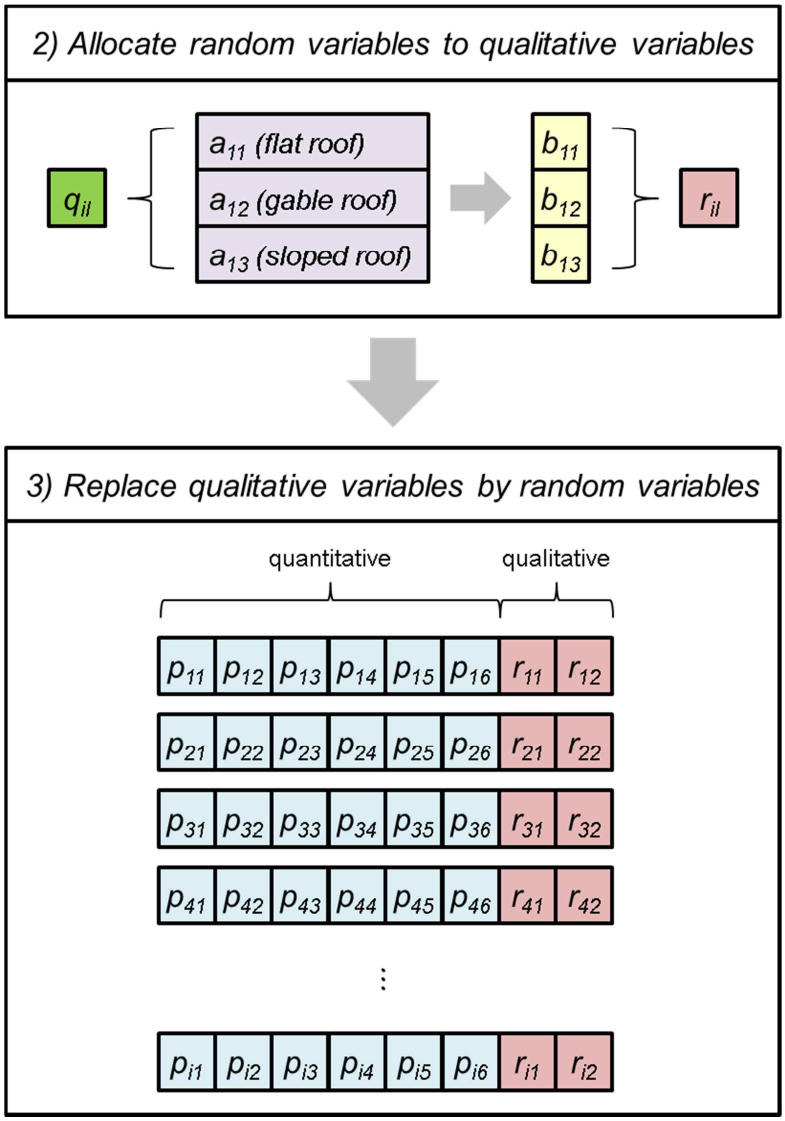


Figure 3-1 Allocating Random Variables to Qualitative Variables

After allocating random variables to qualitative variables, the matrix (Eq. 3-5) is represented as follows:

$$\begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{im} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} + \begin{pmatrix} r_{11} & \cdots & r_{1l} \\ \vdots & \ddots & \vdots \\ r_{i1} & \cdots & r_{il} \end{pmatrix} \times \begin{pmatrix} {w'}_1 \\ \vdots \\ {w'}_l \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_i \end{pmatrix}$$

(Eq. 3-7)

A solution that satisfies the above matrix formula does not exist because it is an insoluble set of equations. In order to solve the matrix with a genetic algorithm, the fitness function *F(x)* is established as the sum of absolute values of the distance of each case as follows:

when

$$\begin{pmatrix} c_1 \\ \vdots \\ c_i \end{pmatrix} - \left\{ \begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{im} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} + \begin{pmatrix} r_{11} & \cdots & r_{1l} \\ \vdots & \ddots & \vdots \\ r_{i1} & \cdots & r_{il} \end{pmatrix} \times \begin{pmatrix} {w'}_1 \\ \vdots \\ {w'}_l \end{pmatrix} \right\}$$

$$= \begin{pmatrix} d_1 \\ \vdots \\ d_i \end{pmatrix},$$

$$min\ F(x) = \sum_{k=1}^{i} |d_k|$$

(Eq. 3-8)

where, d_k is the distance between the cost c_i and the sum of products of attribute values of its weight.

The attribute weights and random variables can be optimized to minimize the result value of the fitness function. Figure 3-2 shows the design of the chromosome to specify the problem to be solved. Its parameters are consisted of the attribute weights and random variables. After optimizing, the weights and the value of random variables are utilized for retrieving similar cases in CBR cost estimating model.

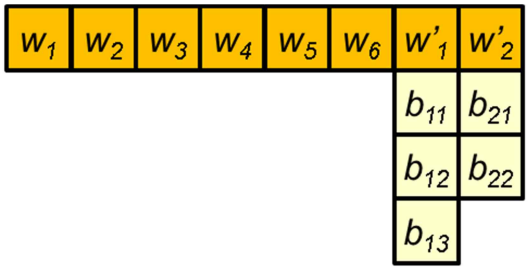


Figure 3-2 Design of Chromosome in GA-CBR Cost Estimating Model

3.1.2 Process of Dual-Optimization

Figure 3-5 shows the process of dual optimization. First, according to the type of variables, quantitative variables are normalized, and qualitative variables are converted into random variables. Next, the initial population consisting of a number of chromosomes is generated randomly, and the chromosomes are evaluated by the fitness function. If the terminal criterion has not been reached, the algorithm is repeated to regenerate chromosomes through crossover and mutation. After the repeated process is finished, the best chromosome which has the lowest value of fitness function is retrieved as the optimal solution from the population.

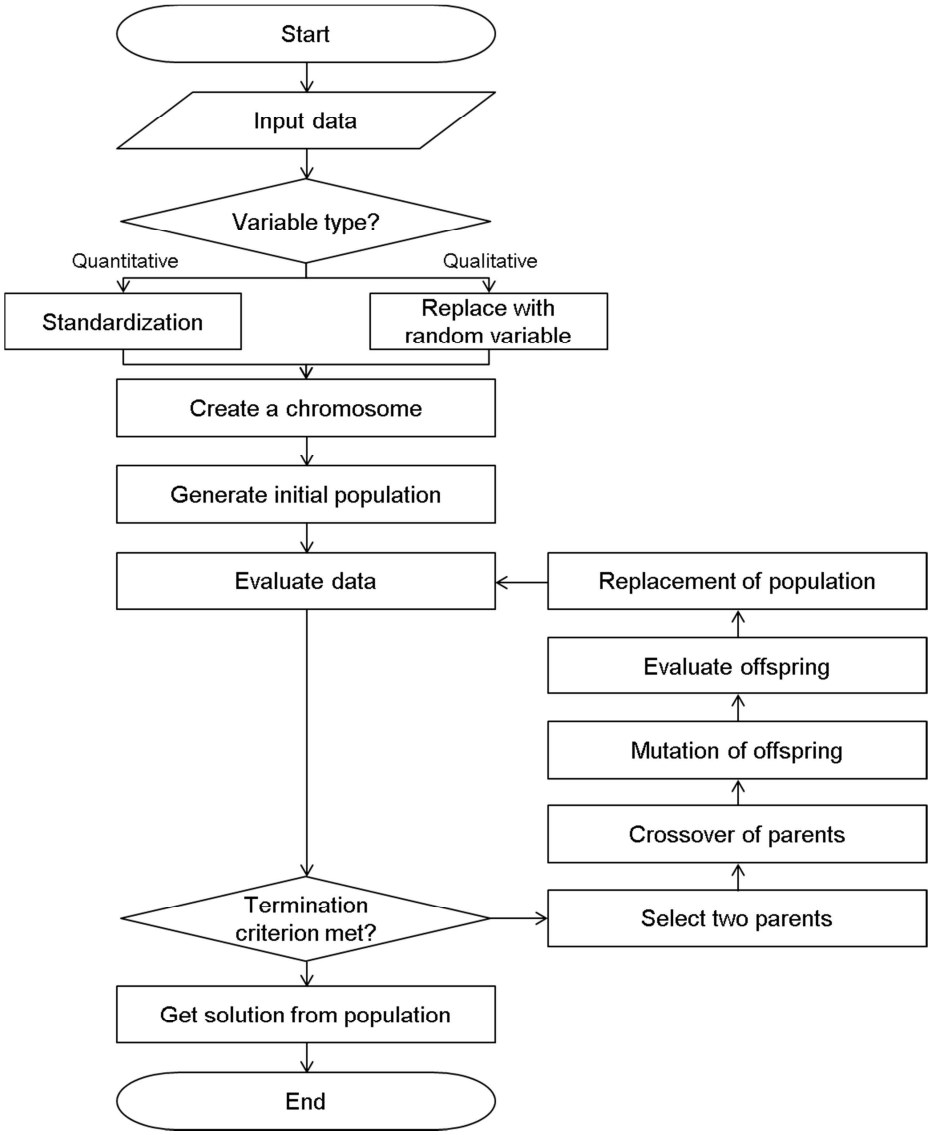


Figure 3-3 Process of Dual-Optimization

3.1.3 Advantages and Disadvantages of Dual-Optimization

Compared to the existing methods, the dual-optimization method presented in this research has the following advantages. First, as above mentioned, it is possible to perform the weight calculation, as well as quantification of the values of qualitative attributes. While previous researches applied the binary method that is not considered the distance between the attributes, or the fuzzy method that is required additional rules or attributes, the dual-optimization method is allocating random variables for qualitative variables, and the weights and the random variables are obtained by optimizing them together. It is difficult to do unless a GA method, which is a global search algorithm.

Secondly, the dual-optimization method is relatively insulated from the problem of attribute selection. In other methods, especially the regression analysis method, the basic premise is that each attribute is independent from each other. However, attributes affecting the construction cost are often not independent from each other. For example, when a value of the attribute 'gross floor area' is increased, the other value of attributes which is related to scale of building such as 'number of households' and 'capacity' tend to increase too. In the AHP method, when other attributes are added to the

existing CBR model, the problem of rank reversal that changes the existing attribute values and their ranking can arise. For instance, when adding the attribute does not affect the construction cost such as 'name of construction manager', the existing attribute values and their ranking might be changed, and also the added attribute affect the construction cost, despite no effect on the construction cost, its weight will be calculated as non-zero value. Because of these problems, existing methods require a lot of effort on the variable selection. However, in the case of dual-optimization, it is free from the multi-collinearity problem because the method is searching the best solution under the given conditions, regardless of whether independently of between variables. Even if the attributes that do not affect the construction cost are included in the cost estimating model, it will present their calculated weights as zero.

Despite the aforementioned advantages, the dual-optimization method has a disadvantage of relatively long calculating time by utilizing more variables than conventional GA methods. Because of the inductive characteristic of finding the optimal solution from the total combinations, GA needs for a sufficiently large number of times of repetition. Also, while the chromosome of previous GA-CBR method is only consisted of the attribute weights, the chromosome of dual-optimization is longer than the previous method because it is consisted of the attribute weights and the

random variables allocated to the qualitative variables. Nevertheless, this disadvantage has been supplemented through the improvement of the algorithm. The weight calculation time of military barracks CBR cost estimating model using dual-optimization is resulted in 44min 43secs at 5,000,000 times, and those of public apartment is resulted in 1hour 3min 45secs at the same number of repetitions. In view of the importance of early stage cost estimation, it is considered that the increase of the calculation time of this degree is enough to be tolerable. Additionally, since the calculated attribute weights is used until the periodic update rather than recalculated each time, it is considered that it does not significantly affect the usability of cost estimation.

3.2 Case Adaptation Methods

In the previous chapter, this research proposes the method for obtaining the attribute weights for retrieving similar cases in the CBR model. Retrieved cases, although their solution can be used directly, should be adjusted to obtain a better solution by case adaptation because there is no case that fits entirely to the problem description. For example, if the problem is to find the construction cost of a building with an area of 100, and the five nearest neighbor cases are retrieved with descending order of similarity, as shown in Table 3-1. The building will be constructed in August 2016, and the retrieved cases were carried out in 2015.

Table 3-1 Examples of Problem and Retrieved Cases

	Area	Distance	Similarity	Cost (year)
Problem	100	-	-	? (2016)
Case 1	95	0.05	0.95	950 (2015)
Case 2	92	0.08	0.92	900 (2015)
Case 3	110	0.10	0.90	1080 (2015)
Case 4	80	0.20	0.80	810 (2015)
Case 5	70	0.30	0.70	700 (2015)

In order to solve the problem, it is needed to consider about how to deal with the following adaptation issues. The first issue is how to adapt the difference between the retrieved similar cases and the problem. Although the retrieved similar cases are closed to the problem, it would be more accurate to use inferring the construction cost when the area is 100, rather than using the original solutions as they are. Like this, identifying the differences between the retrieved cases and the problems, a process of adjusting the solutions of retrieved cases according to the difference affected to the solution is required.

The second issue is how to integrate several retrieved cases to the overall solution. Among retrieved similar cases, there are different case similarities in every case. The more the degree of similarity is higher close to the case in a given problem, which means that the solution of the case is more suitable for solving the problem. If the solution is obtained as simple average value of the solution of retrieved cases without consideration for this issue, it cannot reflect the importance of closer similar case, thereby the result is more likely to become distorted by the solution of case having a relatively low degree of similarity. Therefore, it is necessary that the case closer to the problem has a more effect on deriving the solution by giving the importance according to the degree of similarity of each case.

The last issue is how to convert the cost of solution to the current value. The point of the current problem is in August 2016; however, the time of retrieved cases is 2015. As above mentioned in chapter 2.3.5, this research utilizes the construction cost index for converting the solution to current value.

In order to deal with above three adaptation issues, the following adaptation methods are proposed and applied to the model. Through these methods, this research intends to increase the prediction accuracy of the GA-CBR cost estimating model.

1) The method to adjust retrieving errors of retrieved cases
2) The improved weighted mean method for multiple cases adaptation
3) The method to convert the solution to the current value by using the construction cost index

3.2.1 Retrieving Error Adaptation Method

As above mentioned, the retrieved cases are not entirely consistent with the problem description, so there is a need to adjust the retrieved cases to suit the condition of the problem. This research defines the error caused by the differences between the problem and the similar cases as a retrieving error, and proposes the method to calculate the differences and to adjust the solutions of retrieved cases.

In the GA-CBR cost estimating model, the attribute weight means the relative importance of the attribute. For one attribute, the effect of the difference between the problem and the retrieved case in attribute value on the construction cost can be expressed as the following equation:

$$e_{ij} = w_j(x_{tj} - x_{ij})$$

(Eq. 3-9)

where, e_{ij} is the retrieving error for attribute j of case i, w_j is the weight of attribute j, x_{ti} is the value of j^{th} attribute of a target problem, and x_{ij} is the value of the j^{th} attribute of case i.

When this formula is expanded to all the attributes, it can be described as the following equation:

$$E_i = \sum_{j=1}^{n} e_{ij} = \sum_{j=1}^{n} w_j(x_{tj} - x_{ij})$$

(Eq. 3-10)

where, E_i is the sum of retrieving error of case i.

Thus, the retrieving error which is resulted by difference between a target case and retrieved cases is to be represented as E_i. The cost of retrieved similar cases can be adjusted to the target problem as the following equation.

$$c_i' = c_i + E_i = c_i + \sum_{j=1}^{n} w_j(x_{tj} - x_{ij})$$

(Eq. 3-11)

where c_i' is the adjusted cost of case i.

Since the construction cost c_i is the value converted to the standard normal cumulative distribution in the dual-optimization process, the adjusted cost c_i' should be converted to real value in order to derive a solution. c_i' can be converted to real value through the following equation.

$$When\ F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} exp\ \left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt,$$

$$C_i' = F^{-1}(c_i'|\mu,\sigma)$$

(Eq. 3-12)

where F(x) is the function of normal cumulative distribution, C_i' is the real value of cost of case i, c_i' is the adjusted cost of case i, μ is the average of cost in case base, and σ is the standard deviation of cost in case base.

Figure 3-4 shows the process of adaptation with retrieving error in the GA-CBR cost estimating model. First, the problem information is input to the GA-CBR cost estimating model then similar cases are retrieved. Next, the difference between the problem and retrieved cases are analyzed, and the retrieving errors of retrieved cases are calculated. By adding the error to the solution of each case, the retrieved cases are adjusted to fit to the problem description.

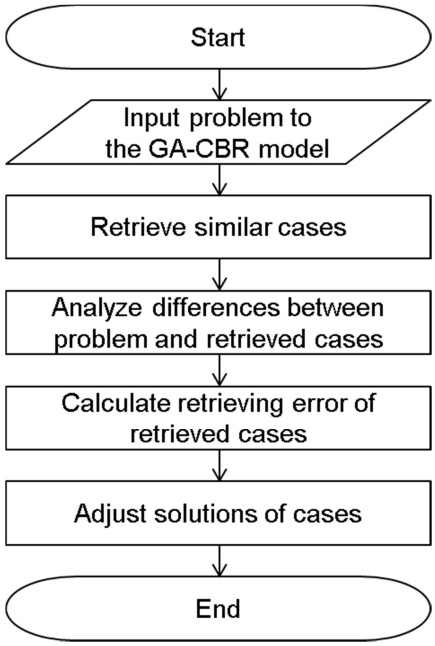


Figure 3-4 Process of Adaptation with Retrieving Error

As above mentioned in chapter 2, the previous adaptation methods for the differences between a problem and similar cases have a disadvantage that requires additional analysis or databases. On the other hand, the retrieving error adaptation method can adapt without these additional elements because it uses the weights obtained in the retrieving error process. This method is limited for qualitative attributes if the relative difference between attribute values of a qualitative attribute. However, this research can quantify the qualitative attribute values by using dual-optimization; therefore, it is possible to apply this method to qualitative attributes.

3.2.2 Improved Weighted Mean Method for Multiple Case Adaptation

There are many methods for multiple case adaptation such as the equal mean, the median, the weighted mean, and the regression analysis. Among the methods for multiple case adaptations, this research uses the weighted mean method which can reflect the difference of case similarities to deduce a solution by giving higher weights to more similar cases. When k is 5, the solution of construction cost can be calculated as following equation.

$$C_s = \sum_{k=1}^{5} w_{kj} \times C_k = \sum_{k=1}^{5} \frac{S_k \times C_k}{\sum S_k}$$

(Eq. 3-13)

where c_s is the construction cost of solution, C_k is the construction cost of retrieved case k, w_{kj} is the weight of retrieved case k, and S_k is the similarity of retrieved case k.

Although the weighted mean method can adapt solutions of retrieved cases based on the degree of case similarity; however, there is a disadvantage that the difference of weights is calculated relatively small. For example, as shown in Table 3-6, when the similarities of retrieved cases

are 0.99, 0.98, 0.97, 0.96, and 0.95, respectively, the ratio of the weight of the most similar retrieved cases to the least similar one is only 1.04 (=0.2041/0.1959). Moreover, as shown in Table 3-2, when the similarities of retrieved cases are 0.99, 0.99, 0.99, 0.99, 0.80, respectively, the ratio of weight is only 1.24 despite of the difference of similarity. This happens very frequently because retrieved cases generally have relatively high similarity values and the difference between weights has decreasing as the value of k increases.

In order to help the problem, this research proposes the improved weighted mean method that utilized the value of standard normal cumulative distribution for the similarity of retrieved cases. This method makes it possible to express the importance of the retrieved cases by utilizing the position relative to the average of the similarity. For the similarities of each case, the value of standard normal cumulative distribution can be calculated as following equation.

$$when\ F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} exp\ \left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt,$$

$$s_k = F(S_k|\mu,\sigma)$$

(Eq. 3-14)

where F(x) is the function of normal cumulative distribution, s_k is the normalized value of similarity of retrieved case k, S_k is the original value of similarity of retrieved case k, μ is the average of similarities of retrieved cases, and σ is the standard deviation of similarities of retrieved cases.

By using the converted values, the solution of construction cost can be calculated as following equation.

$$C_s = \sum_{k=1}^{5} w'_{kj} \times C_k = \sum_{k=1}^{5} \frac{s_k \times C_k}{\sum s_k}$$

(Eq. 3-15)

where Cs is the construction cost of solution, C_k is the construction cost of retrieved case k, w'_{kj} is the weight of retrieved case k, s_k is the normalized value of similarity of retrieved case k, respectively.

As shown in Table 3-2, the ratio of the weight of the most similar retrieved cases to the least similar one is increasing from 1.04 to 8.71(=0.3588/0.0412), from 1.24 to 18.27(=0.2466/0.0135), respectively. By assigning weights according to the similarity distribution of retrieved cases, the improved weighted mean method can increase the influence of more similar cases.

Table 3-2 Comparisons of Multiple Case Adaptation Methods

(a) Example 1

Retrieved Case	Weighted mean		Improved weighted mean	
	Similarity	Weight	Distribution of Similarity	Weight
Case 1	0.99	0.2041	0.8970	0.3588
Case 2	0.98	0.2021	0.7365	0.2946
Case 3	0.97	0.2000	0.5000	0.2000
Case 4	0.96	0.1979	0.2635	0.1054
Case 5	0.95	0.1959	0.1030	0.0412

(b) Example 2

Retrieved Case	Weighted mean		Improved weighted mean	
	Similarity	Weight	Distribution of Similarity	Weight
Case 1	0.99	0.2080	0.6726	0.2466
Case 2	0.99	0.2080	0.6726	0.2466
Case 3	0.99	0.2080	0.6726	0.2466
Case 4	0.99	0.2080	0.6726	0.2466
Case 5	0.80	0.1681	0.0368	0.0135

3.3 Summary

This chapter presents three methods to improve the estimation accuracy of GA-CBR cost estimating model. Firstly, in order to overcome the disadvantage that the existing cost estimating model does not reflect the difference between the values of qualitative attributes, the dual-optimization method is suggested that can calculate the attribute weights as well as quantify the values of qualitative attributes by assigning random variables to each attribute value of the qualitative variable and optimizing these together with the attribute weights by using genetic algorithm. The proposed method has the advantages that it is able to calculate the attribute weights and the values of qualitative attributes at once, and it is relatively free from the attribute selection. However, the computation time is relatively long due to the increase of the chromosome length.

Secondly, this research proposes two new adaptation methods to address the adaptation issues. In order to adjust the errors caused by differences between the problems and similar cases, the retrieving error adaptation method is developed that calculates the retrieving errors using given attribute weights and adjusts the solution based on the calculated retrieving errors. The proposed method has the advantage that it can be calibrated for

differences without additional analysis or databases. It is restricted that can be used only when qualitative attributes are quantified.

In the multiple case adaptation, although the weighted mean method can reflect the difference of case similarities to deduce a solution by giving higher weights to more similar cases based on the case similarity, there is a disadvantage that the difference of weights is calculated relatively small. To solve this issue, this research proposes the improved weighted mean method that utilized the value of standard normal cumulative distribution for the similarity of retrieved cases. By assigning weights according to the similarity distribution of retrieved cases, the improved weighted mean method can increase the influence of more similar cases.

Chapter 4. GA-CBR Cost Estimating Models with Dual-Optimization

Based on the three methods proposed in chapter 3, the GA-CBR cost estimating models for military barracks and public apartment projects are developed. For this, the information of the cases for two projects are collected and established into the database. The costs of cases are normalized by using the construction cost index, and then the attributes that affect the construction cost are derived by literature reviews and expert interviews. The case bases for two cost estimating models are developed based on the derived attributes, and then the attribute weights of the two models are calculated by using the dual-optimization method.

Next, this research presents the cost estimating process using the model. First, when the user enters a problem into the model, five similar cases are retrieved through the nearest neighbor retrieval using the calculated attribute weights. The solutions of the retrieved cases are adapted by the retrieving error adaptation and improved weighted mean methods. Finally, the solution obtained by adaptation is converted to the current value by using the construction cost index.

4.1 Case Base Establishment

A quality of CBR is largely dependent on the structure and content of its case base (Aamodt and Plaza 1994). A case base contains a collection of cases that is used in the context of the CBR methodology for the purpose of performing a reasoning task. In this research, the case base of CBR cost estimating model is established by using two kinds of projects; military barracks projects and public apartment projects in Korea. The process of establishing the case base of CBR cost estimating model involves the following steps:

1) The raw data of two projects are modeled as database with project and building information.
2) From the database, special cases which are out of the ordinary are excluded.
3) After developing the database, the cost data are normalized in terms of escalation by using the construction cost index.
4) From the database, attributes considered to affect the costs were chosen by literature review and interviewing experts.
5) Extracted attributes and costs are re-organized to the case bases in the form of matrices.

.

4.1.1 Data Modeling

The goal of data collection is to utilize historical data to find a solution of the current problem. In order to give the data reliability of the historical data to the user, it should have enough data to ensure statistical significance. This research collects the two kinds of data; Korean military barracks projects ordered by Republic of Korea Ministry of National Defense and Korean public apartment projects conducted by Seoul Housing Corporation, by analyzing their bill of quantities, design drawings, and design description. For efficient data collection and analysis, this research developed the data analysis form for each specific facility by using the MS Excel 2010, and the collected data is stored in the form of database. The database is developed as shown in Table 4-1 that is made up of project cost, general project information, and building information. The database of military barracks has 205 cases with 19 attributes and the database of public apartments has 124 cases with 21 attributes. Excluding special cases similar to build-transfer-lease cases and design-build projects, there were 114 cases of military barracks and 96 cases of public apartment used in this research.

Table 4-1 Data Profile and Database Configuration

	Military Barracks	Public Apartments
Number of total cases	205	124
Used cases	114	96
Year	2005-2008	2006-2008
Location	South Korea	Seoul, South Korea
General project information	name, military type, location, year	name, region, location, year
Building information	capacity, roof shape, structure type, number of floor, number of underground floor, gross floor area, unit floor area, building floor area, quarter area ratio, office area ratio, dining area ratio, bathhouse area ratio, pile foundation, type of heating, type of air conditioning	number of households, gross floor area, number of unit floor households, number of floors, number of underground floors, number of elevators, number of households of unit floor per elevator, number of pilotis with household, shape of pilotis, height of pilotis, height between stories, depth of pit, hallway type, roof type, building shape, structure type, roof terrace

4.1.2 Data Analysis

After developing the database, the cost data need to be normalized in terms of escalation and regional location. This research only considered the escalation effect in this research because Korea's territory is relatively small. The projects were distributed between the years 2005 and 2008; the construction cost values in the database were normalized for the year 2008 using a Korean construction cost index published by the Korea Institute of Civil Engineering and Building Technology (KICT). The converted cost is calculated by multiplying the conversion factor to an existing construction cost as following equation and the conversion factors are shown in Table 4-2.

$$C_{(2008)} = C_{(year)} \times CF_{(year)} = C_{(year)} \times \frac{CI_{(2008)}}{CI_{(year)}}$$

(Eq. 4-1)

where, $C_{(2008)}$ is the converted cost in 2008, $C_{(year)}$ is the original cost, $CF_{(year)}$ is the conversion factor, $CI_{(year)}$ is the cost index in each year, and $CI_{(2008)}$ is the cost index in 2008.

Table 4-2 Conversion Factors for Normalization

Year	306_Non-Housing (Military barracks)		305_Housing (Public apartments)	
	Cost index	Conversion factor	Cost index	Conversion factor
2005	74.15	1.243	76.84	1.207
2006	76.06	1.212	78.43	1.182
2007	78.76	1.171	81.02	1.144
2008	92.19	1.000	92.71	1.000

The collected data are coexisted quantitative and qualitative values and their resulting value is the construction cost. In order to effectively represent the cost, the attributes affecting the cost changes should be extracted. This research uses the same database as that has been used in Ji (2011) and Ahn (2016), and the attribute selection for two cost estimating model is performed in a similar way. First, a pool of attributes considered to affect the costs is developed based on the literature reviews and expert interviews. Then regression analysis is conducted for quantitative attributes to reduce the number of attributes. Finally, the attributes are selected and confirmed by the experts. Seven military officers (Army, Navy, Air Force, and Marine Corps) who had worked as cost estimators over ten years and four construction costs experts who had participated in the construction industry over fifteen years help for selecting attributes of two cost

estimating model. Table 4-3 summarizes the attribute selection process of two cost estimating models.

Table 4-3 Attribute Selection Process

	Military Barracks		Public Apartments	
	Quantitative Attributes	Qualitative Attributes	Quantitative Attributes	Qualitative Attributes
Database	year, capacity, number of floor, number of underground floor, gross floor area, unit floor area, building floor area, quarter area ratio, office area ratio, dining area ratio, bathhouse area ratio	name, military type, location, roof shape, structure type, pile foundation, type of heating, type of air conditioning	year, number of households, gross floor area, number of unit floor households, number of floors, number of underground floors, number of elevators, number of households of unit floor per elevator, number of pilotis with household, height of pilotis, height between stories, depth of pit	name, region, location, shape of pilotis, hallway type, roof type, building shape, structure type, roof terrace
1st screening (Literature reviews and expert interviews)	capacity, number of floor, number of underground floor, gross floor area, unit floor area, building floor area, quarter area ratio, office area ratio, dining area ratio, bathhouse area ratio	military type, roof shape, structure type, pile foundation, type of heating, type of air conditioning	number of households, gross floor area, number of unit floor households, number of floors, number of underground floors, number of elevators, number of households of unit floor per elevator, number of pilotis with household, height of pilotis, height between stories, depth of pit	shape of pilotis, hallway type, roof type, building shape, structure type, roof terrace

2nd screening (Regression analysis)	capacity, number of floor, gross floor area, building floor area, quarter area ratio, office area ratio, dining area ratio, bathhouse area ratio	military type, roof shape, structure type, pile foundation, type of heating, type of air conditioning	number of households, gross floor area, number of unit floor households, number of floors, number of elevators, number of households of unit floor per elevator, number of pilotis with household	shape of pilotis, hallway type, roof type, building shape, structure type, roof terrace
3rd screening (Expert interviews)	capacity, number of floor, gross floor area, building floor area	military type, structure type	number of households, gross floor area, number of unit floor households, number of floors, number of elevators, number of households of unit floor per elevator, number of pilotis with household	hallway type, roof type, building shape

As shown in Table 4-4, in the early stage of the project, six attributes consisted of four quantitative (capacity, number of floors, gross floor area, and building area) and two qualitative (military type and structure type) attributes affecting the construction cost of military barracks are identified. Similarly, as shown in Table 4-5, ten attributes consisted of seven quantitative (number of households, gross floor area, number of unit floor households, number of floors, number of elevators, number of households of unit floor per elevator, and number of pilotis with households) and three qualitative (hallway type, roof type, building shape) attributes are extracted, all of which affect the construction cost of public apartments.

Table 4-4 Attributes used in the GA-CBR Cost Estimating Model (Military Barracks)

Type	Attribute	Unit	Range
Quantitative variables	(P1) Capacity	People	8 – 656
	(P2) Number of floors	floors	1 – 4
	(P3) Gross floor area	m^2	83 – 9,160
	(P4) Building area	m^2	83 – 3,378
Qualitative variables	(Q1) Military type	-	Army, Navy, Air Force
	(Q2) Structure type	-	RC, steel, mixed(RC+steel)

Table 4-5 Attributes used in the GA-CBR Cost Estimating Model (Public Apartments)

Type	Attribute	Unit	Range
Quantitative Variables	(P1) Number of households	ea	5 – 56
	(P2) Gross floor area	m^2	545 – 6,189
	(P3) Number of unit floor households	ea	1 – 6
	(P4) Number of floors	ea	3 – 15
	(P5) Number of elevators	ea	0.5 – 3
	(P6) Number of household of unit floor per elevator	ea	2 – 4
	(P7) Number of pilotis with household	ea	0 – 8
Qualitative Variables	(Q1) Hallway type	-	corridor type, stairway type
	(Q2) Roof type	-	flat roof, gable roof, slope roof
	(Q3) Building shape	-	linear shape, L shape

After modeling and analyzing the data, extracted attributes and converted costs are re-organized to the case bases in the form of matrix such as its presented in Figure 4-1. The cases are represented in rows and the attributes and cost are represented in columns. The attribute values and costs are represented where X_{ij} represent the j^{th} attribute value of case i and C_i represent the converted cost of case i. Case base of military barrack

projects are shown in Appendix 2, and those of public apartment projects are shown in Appendix 3.

Case No.	Attributes					Cost (2008)
	1	2	3	...	j	
Case 1	X_{11}	X_{12}	X_{13}	...	X_{1j}	C_1
Case 2	X_{21}	X_{22}	X_{23}		X_{2j}	C_2
Case 3	X_{31}	X_{32}	X_{33}		X_{3j}	C_3
⋮	⋮				⋮	⋮
Case i	X_{i1}	X_{i2}	X_{i3}	...	X_{ij}	C_i

Figure 4-1 Matrix of Case Base

4.2 Calculating Weights using Dual-Optimization

4.2.1 Model 1: Military Barrack Projects

In order to calculate the attribute weights, the values of four quantitative attributes and cost in the case base should be normalized. For this, the values of mean and standard deviation for attributes and cost should be obtained first. The calculated mean and standard deviation values for military barracks projects cost estimating model is shown in Table 4-6.

Table 4-6 Summary of Mean and Standard Deviation

	Mean	Standard Deviation
Capacity	110.03	98.86
Number of floors	2.13	0.84
Gross floor area	2,054.44	1927.19
Building area	905.31	604.39
Cost (2008)	1,883,089,559	1,607,041,600

Based on the calculated values of mean and standard deviation, the values of quantitative attributes and cost are converted to the value of standard normal cumulative distribution by using the function of

'NORM.DIST' in the program of MS Excel 2010. This allows each value to be converted to a value between 0 and 1. Since this method cannot be applied to qualitative attributes, random variables having a value between 0 and 1 are assigned to each value of two qualitative attributes. Through this process, the attributes and cost values are represented by a value of 0 to 1.

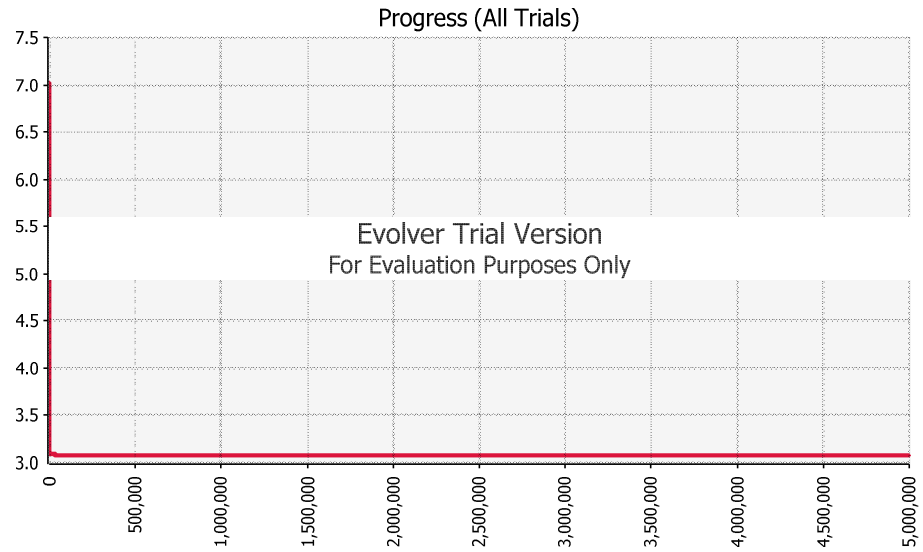
Based on these data, the dual-optimization method is performed to calculate the values of attribute weights and random variables. The optimizations of these values were conducted by using commercial GA software to add-on for the MS Excel, 'Evolver 7.5' (Palisade, http://www.palisade.com/evolver/). The conditions for the GAs were a crossover rate of 0.05, a mutation rate of 0.1, and an initial population of 50. These are the default values by the program 'Evolver 7.5'. In addition, the termination criterion was met when the number of trials reached 5,000,000. In the MS Excel, six attribute weights and six random variables (three of Military type and three of Structure type) were assigned as adjustable cells, and sum of distance, which is sum of the distance of each case, was assigned as optimization cell. For adjustable cells, the recipe solving method, which is the simplest and most frequently used solving method, was used for solving method that the adjustable cells can be varied independently of one another. The type of all adjustable cells is set to 'any', which means that all real number are included. The optimization goal is

settled to minimum and all sub-operators provided by the program were set to on. Table 4-7 summarizes the sub-operators used for optimization.

Table 4-7 Sub-operators in Evolver Program

Sub-operator	On/off
Default parent selection	On
Default mutation	On
Default crossover	On
Default backtrack	On
Heuristic crossover	On
Cauchy mutation	On
Boundary mutation	On
Non-uniform mutation	On
Linear	On
Local search	On

The summary of optimization is shown in Figure 4-2. The sum of distance started at 16.2548 for the first time and it could be seen to converge quickly to 3.0838. As a result of the optimization, the value of the 4,898,800th calculation was selected to the best value. The time to find best value was 43min 18sec and the total optimization time was 44min 43sec.



Goal	
Cell to Optimize	'GAopt(Dual)'!Z9
Type of Goal	Minimum

Results	
Valid Trials	5000000
Total Trials	5000000
Original Value	16.25484731
+ soft constraint penalties	0.00
= result	16.25484731
Best Value Found	3.083819284
+ soft constraint penalties	0.00
= result	3.083819284
Best Trial Number	4898800
Time to Find Best Value	0:43:18
Reason Optimization Stopped	Number of trials
Time Optimization Started	2016-11-08 7:02
Time Optimization Finished	2016-11-08 7:47
Total Optimization Time	0:44:43

Figure 4-2 Summary of Optimization (Military Barrack Projects)

The optimized values for adjustable cells were summarized as Table 4-8. For the qualitative attributes, their attribute weights and the values of random variables were readjusted so that the maximum value of a random variable was 1.0000. According to the result of optimization, the attribute of Gross floor area was the highest value of 0.9152 among all the attribute weights and next came the military type (0.0283), the number of floors (0.0274), the building area (0.0215), the capacity (0.0151), and the structure type (0.0090). This means that the attribute of Gloss floor area has the largest effect on the construction cost.

In the military attribute, the navy's quantifying value was relatively high at 1.0000, compared to the air force's value of 0.2289 and the army's value of 0.0000. This means that the navy's barracks are more expensive when the exactly same type of building is built. In the structure attribute, RC was recorded as the highest quantified value at 1.0000 compared to steel's value of 0.9497 and that of mixed structure of 0.0000.

Table 4-8 Result of Optimization (Military Barrack Projects)

Attribute	Weight	Random variable		Value
(P1) Capacity	0.0151	(Q1) Military type	Army	0.0000
(P2) Number of floors	0.0274		Navy	1.0000
(P3) Gross floor area	0.9152		Air Force	0.2289
(P4) Building area	0.0215	(Q2) Structure type	RC	1.0000
(Q1) Military type	0.0283		Steel	0.9497
(Q2) Structure type	0.0090		Mixed	0.0000

4.2.2 Model 2: Public Apartment Projects

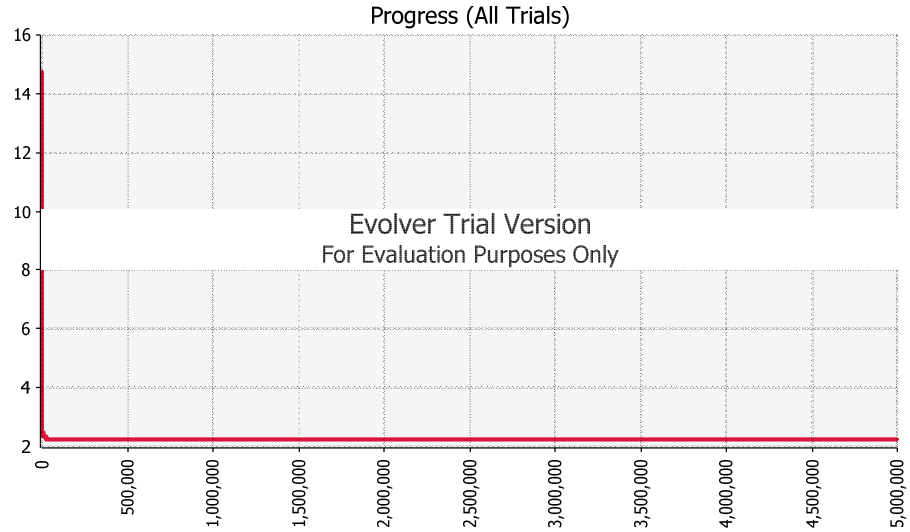
For public apartment projects cost estimating model, calculating the attribute weights and quantifying the qualitative attributes are conducted in a similar way to the model of military barracks projects by using the dual-optimization method. To calculate the attribute weights, the values of seven quantitative attributes and cost in the case base should be normalized. For this, the values of mean and standard deviation for attributes and cost should be obtained first. The calculated mean and standard deviation values for military barracks projects cost estimating model is shown in Table 4-9.

Table 4-9 Summary of Mean and Standard Deviation

	Mean	Standard Deviation
Number of households	27.80	15.52
Gross floor area	3,059.60	1,716.83
Number of unit floor households	2,83	1.16
Number of floors	10.16	3.38
Number of elevators	1.14	0.53
Number of household of unit floor per elevator	2.64	0.90
Number of pilotis with households	1.55	1.78
Cost (2008)	1,535,823,739	892,792,276

Based on the calculated values of mean and standard deviation, the values of quantitative attributes and cost were converted to the value of standard normal cumulative distribution, and random variables having a value between 0 and 1 are assigned to each value of three qualitative attributes. In the same way as the above model, the values of attribute weights and random variables were optimized by using the dual-optimization method. Ten attribute weights and seven random variables (two of Hallway type, three of Roof type, and two of Building shape) are assigned as adjustable cells, and sum of distance, which is sum of the distance of each case, was assigned as optimization cell.

The summary of optimization is shown in Figure 4-3. The sum of distance started at 14.7786 for the first time and it could be seen to converge quickly to 2.2204. As a result of the optimization, the value of the 4,898,800th calculation was selected to the best value. The time to find best value was 1hour 3min 8sec and the total optimization time was 1hour 3min 45sec.



Goal	
Cell to Optimize	'GAopt(Dual)'!AL9
Type of Goal	Minimum

Results	
Valid Trials	5000000
Total Trials	5000000
Original Value	14.77860525
+ soft constraint penalties	0.00
= result	14.77860525
Best Value Found	2.220380433
+ soft constraint penalties	0.00
= result	2.220380433
Best Trial Number	4996547
Time to Find Best Value	1:03:08
Reason Optimization Stopped	Number of trials
Time Optimization Started	2016-11-09 6:20
Time Optimization Finished	2016-11-09 7:24
Total Optimization Time	1:03:45

Figure 4-3 Summary of Optimization (Public Apartment Projects)

The optimized values for adjustable cells were summarized as Table 4-10. For the qualitative attributes, their attribute weights and the values of random variables were readjusted so that the maximum value of a random variable was 1.0000. According to the result of optimization, the attribute of Gross floor area was the highest value of 0.7177 among all the attribute weights and next came the number of households (0.1455), the hallway type (0.1116), the roof type (0.0920), the number of pilotis with households (0.0223), the number of elevators (0.0155), and the building shape (0.0041). Three attributes (number of unit floor households, number of floors, and number of household of unit floor per elevator) were resulted in 0.0000.

In the hallway type, the corridor type was relatively high at 1.0000 compared to the stairway type's value of 0.0000. In the roof type, the slope roof was recorded as the highest quantified value at 1.0000 compared to the gable roof's value of 0.3127 and that of the flat roof of 0.0000. In the building shape, the L shape was relatively high at 1.0000 compared to the linear shapes' value of 0.0000.

Table 4-10 Result of Optimization (Public Apartment Projects)

Attribute	Weight	Random variable		Value
(P1) Number of households	0.1455	(Q1) Hallway type	Corridor	1.0000
(P2) Gross floor area	0.7177		Stairway	0.0000
(P3) Number of unit floor households	0.0000	(Q2) Roof type	Flat	0.0000
(P4) Number of floors	0.0000		Gable	0.3127
(P5) Number of elevators	0.0155		Slope	1.0000
(P6) Number of household of unit floor per elevator	0.0000	(Q3) Building shape	Linear shape	0.0000
(P7) Number of pilotis with household	0.0223		L shape	1.0000
(Q1) Hallway type	0.1116			
(Q2) Roof type	0.0920			
(Q3) Building shape	0.0041			

4.3 Cost Estimating Model

4.3.1 System Architecture

The system architecture of the GA-CBR cost estimating model is shown in Figure 4-4. First, user input attributes of a problem into the cost estimating model. Then the similarities of all case in the case base are calculated based on the attribute weights obtained by the dual-optimization. Cases are sorted according to their similarity in descending order, and then the top five cases are retrieved as similar cases. For the retrieved cases, differences between problem and retrieved cases are analyzed and their retrieving errors are calculated. These are used to adjust solutions of the retrieved cases. The case weights are calculated based on the distribution of their similarity, and then the improved weighted mean of solution is obtained. Finally, the solution is converted to current value by the construction cost index.

The obtained solution is applied to solve problem. If the solution proves to be failed to solve the problem, the solution will be revised. After the solution has been successfully adapted to the target problem, store the resulting experience as a new case in memory. The more CBR is used, the

more case are stored in case base. When the number of stored cases reaches a certain level or more, new weights should be calculated by using the dual-optimization.

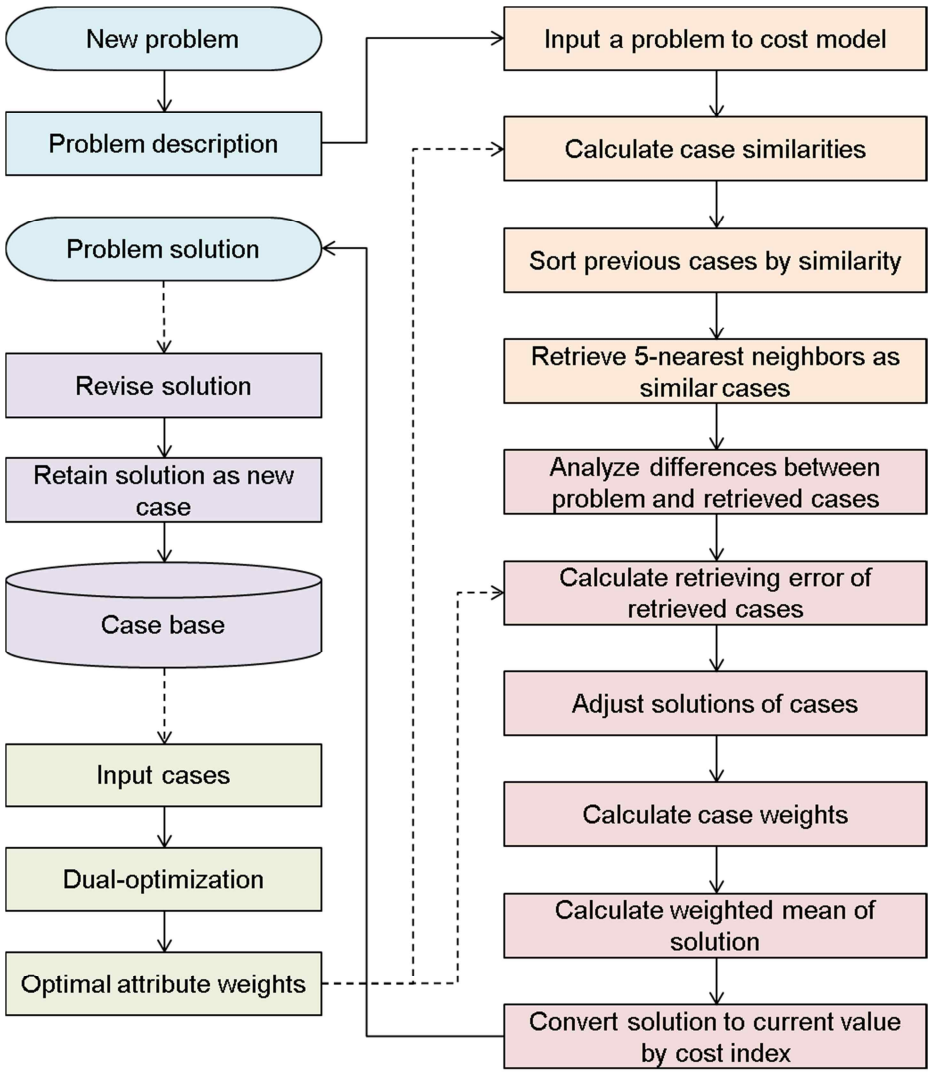


Figure 4-4 System Architecture of the GA-CBR Cost Estimating Model

4.3.2 Cost Estimating Process

The following is an example of cost estimation for military barracks project. When a problem is given that finding the construction cost of building at August 2016 as shown in Table 4-11, enter the values of 6 attributes of the problem into the cost estimating model. The values of quantitative attribute are converted to the normalized values, and those of qualitative attribute are changed to quantified values obtained by the dual-optimization.

Table 4-11 Problem Description

Problem (August 2016)	P1	P2	P3	P4	Q1	Q2
Original value	30	2	628.4	322.2	Army	Steel
Converted value	0.2091	0.4372	0.2297	0.1673	0.0000	0.9316

Using the information, the CBR model performs calculating similarities of all case for retrieving similar cases. Based on the case similarities, the top five cases are extracted with high similarity order. Table 4-12 summarizes the retrieved five similar cases in the GA-CBR cost estimating model.

Table 4-12 Retrieved cases

	P1	P2	P3	P4	Q1	Q2	Normalized cost (2008)
Case 85	0.2394	0.4372	0.228	0.1642	0	0.9316	0.2423
Case 71	0.2239	0.4372	0.2299	0.2267	0	0.9316	0.266
Case 46	0.2239	0.4372	0.222	0.1541	0	0.9809	0.2053
Case 77	0.1607	0.4372	0.2175	0.1439	0	0.9809	0.2171
Case 57	0.2331	0.4372	0.2453	0.1818	0	0.9316	0.2241

In the GA-CBR cost estimating model, the retrieving error of each case is calculated as sum of the attribute weight multiplied by the difference of attribute between the problem and the retrieved case. Then it is added to the value of normalized cost. Since these calculated values are represented in the standard normal cumulative distribution, these are reconverted to the original unit. This process was performed by using the function of 'NORM.INV' in the MS Excel. Table 4-13 summarizes the results of retrieving error adaptation.

Table 4-13 Results of Retrieving Error Adaptation

Retrieved case	Normalized cost (2008)	Retrieving error	Adjusted normalized cost (2008)	Adjusted cost (2008)
Case 85	0.2423	0.0011	0.2433	₩ 765,210,336
Case 71	0.2660	-0.0017	0.2643	₩ 870,491,406
Case 46	0.2053	0.0065	0.2118	₩ 597,203,974
Case 77	0.2171	0.0116	0.2287	₩ 689,034,136
Case 57	0.2241	-0.0146	0.2095	₩ 584,310,168

The derived five adjusted costs are converted to one solution through the improved weighted mean method. Based on the similarities of retrieved cases, the case weights are calculated by using equation 3-14 and 3-15. These are multiplied by the adjusted costs and sum of them is derived to one solution, as shown in Table 4-14.

Table 4-14 Results of Multiple Cases Adaptation

Retrieved case	Adjusted cost (2008)	Similarity	Normalized similarity	Case weight	Solution (2008)
Case 85	₩ 765,210,336	99.60%	0.9155	0.3699	₩ 734,738,944
Case 71	₩ 870,491,406	99.12%	0.6238	0.2520	
Case 46	₩ 597,203,974	99.10%	0.6041	0.2441	
Case 77	₩ 689,034,136	98.59%	0.1947	0.0787	
Case 57	₩ 584,310,168	98.48%	0.1369	0.0553	

Since the derived solution is based on 2008, it is converted to current value by using the construction cost index, as shown in Table 4-15. The solution is multiplied by the conversion factor to derive the estimated cost. Finally, the construction cost is estimated to ₩ 966,411,101.

Table 4-15 Example of Cost Conversion to Current Value

Solution (2008)	Cost index of 2008	Cost index of August 2016	Conversion factor	Estimated cost (August 2016)
₩ 734,738,944	92.19	116.43	1.263	₩ 966,411,101

4.4 Summary

In chapter 4, this research develops two GA-CBR cost estimating models for military barracks and public apartment projects based on three methods proposed in chapter 3. To do this, the case bases of two models are established first. The information of the cases are collected and analyzed, and then the attributes that affect the construction cost are derived by attribute selection. Next, the dual-optimization is performed for each model to get the attribute weights and the quantified values of qualitative attributes. Based on the obtained values, the GA-CBR cost estimating models are developed and the cost estimating process is explained in detail. For the given problem, similar cases are retrieved from the case base by using the nearest neighbor retrieval with the obtained attribute weights. The solutions of retrieved cases are adapted by two adaptation methods, the retrieving error adaptation and improved weighted mean methods. The derived solution is converted to the current value by using the construction cost index. Finally, the solution is retained in the case base as a new case.

Chapter 5. Model Validations

In chapter 5, the suggested GA-CBR cost estimating model is validated for military barracks and public apartment projects. The case studies are conducted to examine the accuracy of cost estimation results of the proposed GA-CBR cost estimating model by 10-fold cross validation. In order to validate the performance of the GA-CBR cost estimating model, three validations were performed respectively. In the validation 1, the retrieving accuracy of the GA-CBR cost estimating model using dual-optimization method is verified by comparing to those of cost estimating models using other methods. In the validation 2, the applicability of the retrieving error adaptation method is examined. For the cost estimating model using the dual-optimization, it is verified how the accuracy improves when the adaptation is applied. In the validation 3, the applicability of the improved weighted mean method is validated. For the result of validation 2, it is examined how the result changes when the improved method is applied by comparing the result of applying the weighted mean method.

5.1 Validation Methods

The validation methods for case studies are summarized in Table 4-1. Three validations were conducted by using the two case bases: military barracks and public apartment projects. The nearest neighbor retrieval method was used to retrieve similar cases and five cases were retrieved as similar cases using the weighted Euclidean distance based similarity measurement. Case bases of two projects were normalized by converting to the standard normal cumulative distribution. In validation 1, the attribute weights were calculated by four methods (Dual-optimization, Feature counting, Regression analysis, and GA only quantitative attributes) and each of these was used to retrieve five similar cases. The retrieved cases were reused to estimate the average costs and estimation performances of four cost estimating models were evaluated and compared in terms of estimation accuracy. In validation 2, the cost estimating model using dual-optimization was used to retrieve similar cases and these were adapted by the adaptation of retrieving errors. As similar to validation 1, the average cost of retrieved and adapted cases were used for validation. In validation 3, the estimation results of three multiple case adaptation methods (Mean, Weighted mean, and Improved weighted mean) were evaluated and compared. .

Table 5-1 Validation Methods for Case Studies

	Validation 1 Dual-optimization	Validation 2 Retrieving error adaptation	Validation 3 Improved weighted mean method
Case base	· Military barrack projects · Public apartment projects		
Retrieving method	· 5-Nearest neighbor retrieval		
Normalization method	· Convert to standard normal cumulative distribution		
Similarity measurement	· Weighted Euclidean distance		
Weight calculation method	· Dual-optimization · Feature counting · Regression analysis · GA only quantitative attributes	· Dual-optimization	· Dual-optimization
Retrieving error adaptation	X	O	O
Multiple cases adaptation	· Mean	· Mean	· Mean · Weighted mean · Improved weighted mean
Sampling method	· 10-fold cross validation		
Performance measurement	· MAER		

While performing the machine learning, there are two types of data set: a training set and a test set. It does not matter if there is a sufficient amount of data in the data set, but in actual situations the number of samples cannot be provided infinitely. Therefore, when the amount of data is not sufficient, a resampling technique is used to increase the statistical reliability of model performance measurements. To do this, the most commonly used method is k-fold cross validation method (Seni and Elder 2010, Fushiki 2011). This method divides the samples into k equal-sized groups, trains the model with k-1 sets, and tests the other one set to measure the performance of the model. For all k sub-sets, this process is repeated k times and the average of accuracy is defined as the performance of the model. As the k closes to the total number of samples, most of samples can be used for training, but that takes a lot of time. Although the k is an unfixed parameter, 10-fold cross validation is commonly used (McLachlan et al. 2005). Figure 5-1 shows the k-fold cross validation.

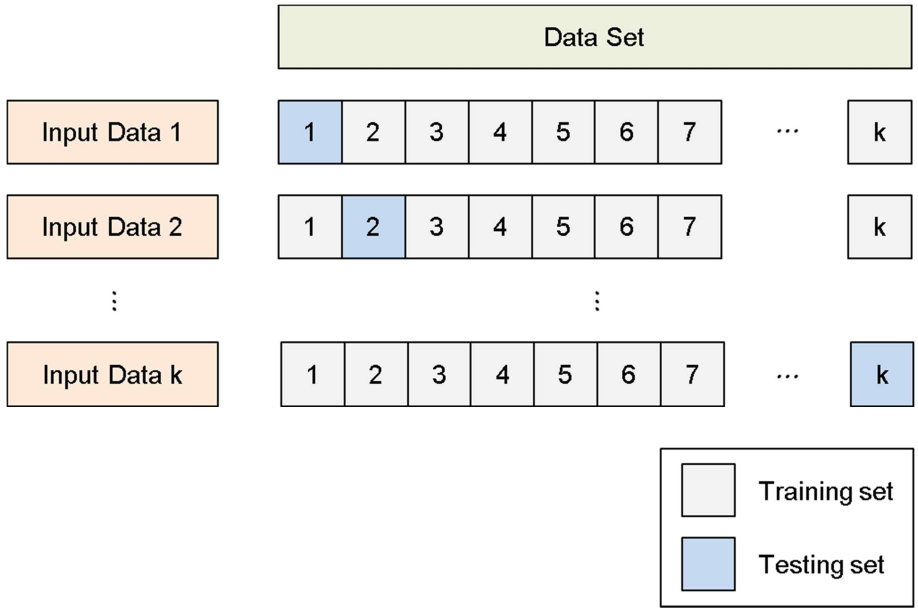


Figure 5-1 k-fold Cross Validation

For each validation case, the accuracy of cost estimation was measured by the absolute error ratio (AER) represented in Equation 5-1. It is an indicator of how close the estimated cost is to the actual cost. The estimation accuracy for each fold of test set was measured by the mean of absolute error ratios (MAER), as shown in Equation 5-2. Comparing the average of MAER allows the determination of which methods are more accurate. If the MAER is closer to 0, it shows that the estimation accuracy of model is higher.

$$AER(\%) = \frac{\left|C_t - \hat{C}_t\right|}{\hat{C}_t} \times 100\ (\%)$$

(Eq. 5-1)

$$MAER(\%) = \frac{\sum_{t=1}^{n} \frac{\left|C_t - \hat{C}_t\right|}{\hat{C}_t}}{n} \times 100\ (\%)$$

(Eq. 5-2)

where, C_{kt} is the estimated cost of validation case t, $\hat{C}_{kt}$ is the actual cost of validation case t, and n is the number of validation cases.

In order to derive sub data sets, random numbers were assigned to each case by using the 'RAND' function in the MS Excel 2010. After sorting the cases in ascending order of the random numbers, then the sub data sets were extracted sequentially by using sampling without replacement. In the military barracks projects cost estimating model, there were 114 cases in the case base, so 11 cases were extracted for each fold, and the remaining 4 cases were excluded. In the public apartment projects cost estimating model, there are 96 cases in the case base, so 9 cases were extracted for each fold, and the remaining 6 cases were excluded. Table 5-2 and 5-3 summarize the result of extracting sub data sets from the case base.

Table 5-2 Sub Data Sets (Military Barrack Projects)

Fold No.	Case No.										
1	114	64	55	96	22	34	29	98	106	49	75
2	26	60	84	14	31	93	91	103	13	65	59
3	11	2	110	87	113	109	28	58	1	9	4
4	86	57	15	102	5	23	42	30	61	89	73
5	79	8	38	63	72	81	6	18	52	48	85
6	51	37	27	100	17	82	44	80	67	56	74
7	104	101	66	78	111	95	45	25	41	43	24
8	19	107	7	32	46	92	3	21	39	70	71
9	16	105	20	112	77	97	50	88	83	40	108
10	53	54	68	99	12	62	35	76	94	90	36
remaining	33	47	10	69							

Table 5-3 Sub Data Sets (Public Apartment Projects)

Fold No.	Case No.								
1	73	60	21	84	77	1	76	32	62
2	44	12	3	67	15	2	54	92	46
3	65	70	17	63	57	11	42	5	53
4	38	16	39	28	19	7	24	35	4
5	41	45	30	55	83	56	26	87	96
6	58	78	29	75	52	18	61	10	20
7	85	8	59	49	51	93	27	13	88
8	86	31	40	72	23	80	68	64	79
9	25	48	94	50	95	14	37	74	90
10	91	9	22	6	71	33	66	69	89
remaining	43	34	81	47	36	82			

5.2 Validation 1: Dual-optimization Method

Validation 1 was performed to verify the following hypothesis.

- Hypothesis 1: For the calculating of the attribute weights in the CBR cost estimating model, the estimation accuracy of the dual-optimization method will be superior to other existing methods.

In order to validate the performance of the dual-optimization method, this research conducted case studies using military barracks and public apartment projects by comparing CBR outputs of the dual-optimization method to those of previous weight assignment methods: feature counting, regression analysis, and GA method using only quantitative attributes. For each fold of sub data set, the remaining data were utilized as a training set, and attribute weights of each fold were calculated by four methods and these were used to develop CBR cost estimating models. The cost of each case was estimated by using four CBR cost estimating models respectively. By comparing the average of the MAERs, the effectiveness of the dual-optimization method was compared to others in terms of estimation accuracy.

5.2.1 Military Barracks Projects

1) Calculating Attribute Weights

Based on the case base of military barracks projects, the attribute weights were calculated by using four methods (Dual-optimization, Feature counting, Regression analysis, and GA only quantitative attributes). In the cost estimating model of dual-optimization, the attribute weights were calculated by GA program 'Evolver' in a similar way to chapter 4.2.1. The conditions for the GAs were a crossover rate of 0.05, a mutation rate of 0.1, an initial population of 50, and termination criterion of 5,000,000 trials reached. Six attribute weights and six random variables were assigned as adjustable cells, and sum of distance was assigned as optimization cell. This process was performed for each fold, and a total of 10 attribute weights sets were derived. The attribute weights and random variables calculation results of dual-optimization are summarized in Table 5-4.

Table 5-4 Calculation Results of Dual-Optimization (Military Barrack Projects)

Fold No.	Attribute weight						Value of random variables for qualitative attributes					
	P1	P2	P3	P4	Q1	Q2	Q11	Q12	Q13	Q21	Q22	Q23
1	0.0059	0.0223	0.9344	0.0178	0.0293	0.0113	0.0000	1.0000	0.3620	0.7635	1.0000	0.0000
2	0.0335	0.0524	0.8503	0.0505	0.0279	0.0065	0.0000	1.0000	0.3554	1.0000	0.6806	0.0000
3	0.0123	0.0248	0.9340	0.0164	0.0307	0.0080	0.0000	1.0000	0.7068	0.0132	1.0000	0.0000
4	0.0689	0.0359	0.8779	0.0079	0.1147	0.0062	0.0000	1.0000	0.1049	0.8555	1.0000	0.0000
5	0.0110	0.0172	0.9300	0.0179	0.0271	0.0081	0.0000	1.0000	0.3425	1.0000	0.7842	0.0000
6	0.0047	0.0277	0.9326	0.0188	0.0326	0.0089	0.0000	1.0000	0.3907	0.1345	1.0000	0.0000
7	0.0030	0.0418	0.9192	0.0248	0.0318	0.0063	0.0000	1.0000	0.7565	0.0000	1.0000	0.0000
8	0.0653	0.0466	0.8235	0.0423	0.0221	0.0205	0.0000	1.0000	0.0000	1.0000	0.2668	0.0000
9	0.0439	0.0174	0.9168	0.0054	0.0298	0.0089	0.0000	1.0000	0.4894	1.0000	0.7620	0.0000
10	0.0101	0.0235	0.9326	0.0162	0.0306	0.0087	0.0000	1.0000	0.6361	0.5257	1.0000	0.0000

※ P1: Capacity, P2: Number of floors, P3: Gross floor area, P4: Building area, Q1: Military type, Q2: Structure type, Q11: Army, Q12: Navy, Q13: Air force, Q21: RC, Q22: Steel, Q23: Mixed.

In the cost estimating model of feature counting, the weights of six attributes were set to the same value of 0.1667 regardless of the fold. For the qualitative attributes, the binary method was applied to represent the distance between attribute values. The attribute weights of feature counting are summarized in Table 5-5.

Table 5-5 Attribute Weights of Feature Counting (Military Barrack Projects)

Attribute	P1	P2	P3	P4	Q1	Q2
Weight	0.1667					

In order to calculate the attribute weights of regression analysis cost estimating model, regression analysis was performed for training set of each fold by setting the cost as a dependent variable and the four quantitative variables as independent variables. The absolute value of the regression coefficient obtained through the analysis was used as the weight. The analysis was conducted by the program of MS Excel 2010 with confidence level of 95%. The attribute weights calculation results of regression analysis are summarized in Table 5-6.

Table 5-6 Calculation Results of Regression Analysis (Military Barrack Projects)

Fold No.	P1	P2	P3	P4
1	0.0176	0.0900	0.6979	0.1938
2	0.0190	0.0972	0.6756	0.2088
3	0.0348	0.0897	0.6594	0.2141
4	0.0269	0.0917	0.7017	0.1829
5	0.0399	0.0551	0.8614	0.1077
6	0.0313	0.0748	0.7447	0.1523
7	0.0278	0.1158	0.7050	0.2136
8	0.0726	0.0998	0.6188	0.2045
9	0.0196	0.0863	0.7136	0.1763
10	0.0201	0.0770	0.7094	0.1898

In the cost estimating model of GA with only quantitative attributes, the weights of four quantitative attributes were calculated by GA program 'Evolver' as similar to the dual-optimization. The conditions for the GAs were a crossover rate of 0.05, a mutation rate of 0.1, an initial population of 50, and termination criterion of 5,000,000 trials reached. Four quantitative attribute weights were assigned as adjustable cells, and sum of distance was assigned as optimization cell. The attribute weights calculation results of GA only quantitative attributes are summarized in Table 5-7.

Table 5-7 Calculation Results of GA only Quantitative Attributes (Military Barrack Projects)

Fold No.	P1	P2	P3	P4
1	0.0000	0.0564	0.8818	0.0627
2	0.0515	0.0691	0.8235	0.0583
3	0.0586	0.0512	0.8152	0.0753
4	0.0787	0.0618	0.8307	0.0322
5	0.0000	0.0570	0.8787	0.0654
6	0.0664	0.0310	0.8847	0.0169
7	0.0000	0.0631	0.8792	0.0607
8	0.0898	0.0374	0.8559	0.0197
9	0.0793	0.0422	0.8554	0.0262
10	0.0545	0.0509	0.8476	0.0464

2) Results and Discussions

Based on the above attribute weights, 10-fold cross validation was performed. Forty CBR cost estimating models were developed and the construction costs of validation cases were estimated. Table 5-8 shows the comparison of MAER by four weight calculation methods. In terms of estimation accuracy, although there were small differences in individual test sets, the dual-optimization cost estimating model resulted in the lowest

MAER at 10.15%, compared to feature counting at 22.57%, regression analysis at 12.18%, and GA with only quantitative attributes at 11.07%. In the folds of 4, 5, 6, 8, 9, and 10, the dual-optimization method resulted in lowest MAER compared to the other methods. In the results of maximum MAER, the dual-optimization method resulted in the lowest value of 14.96% compared to feature counting at 22.57%, regression analysis at 18.07%, and GA with only quantitative attributes at 15.88%. Also, the dual-optimization method resulted in the lowest minimum MAER at 5.54%, compared to feature counting at 7.14%, regression analysis at 6.98%, and GA with only quantitative attributes at 6.92%.

More precisely, for the expected accuracy ranges of class 4 and 5 in the Cost Estimate Classification Matrix from AACE, it was analyzed that which number of cases are outside the range. Table 5-9 shows the number of cases beyond the expected accuracy range by four weights calculation methods. For the range from +100 to -50%, which is the maximum expected accuracy range at Class 5, there was no case beyond that range. For the range from +50 to -30%, which is the maximum expected accuracy range at Class 4, the dual-optimization method resulted that all cases were within the range although other methods resulted that there were some cases out of range. In the results of other minimum ranges at Class 5 and 4, the dual-optimization method resulted in the lowest number of cases beyond the expected

accuracy ranges compared to other methods. Collectively, the dual-optimization cost estimating model had best estimation accuracy compared to the cost estimating models of other methods.

Table 5-8 Comparison of MAERs by Weight Calculation Methods (Military Barracks Projects)

(MAER, %)

Fold No.	**Dual-optimization**	Feature Counting	Regression Analysis	GA only quantitative attributes
1	**12.45**	16.68	11.76	11.59
2	**14.96**	18.95	14.84	15.88
3	**10.26**	10.44	13.08	7.94
4	**11.36**	22.57	18.07	12.93
5	**10.15**	14.98	12.00	11.05
6	**8.09**	10.10	12.00	11.99
7	**13.15**	13.44	12.68	13.05
8	**7.39**	16.75	10.99	9.50
9	**8.12**	12.68	9.42	9.89
10	**5.54**	7.14	6.98	6.92
Average MAER	**10.15**	14.37	12.18	11.07
Max MAER	**14.96**	22.57	18.07	15.88
Min MAER	**5.54**	7.14	6.98	6.92

Table 5-9 Number of Cases beyond Expected Accuracy Range by Weight Calculation Methods (Military Barracks Projects)

Expected accuracy range	**Dual-optimization**	Feature Counting	Regression Analysis	GA only quantitative attributes
+100% to -50% (Maximum at Class 5)	**0**	0	0	0
+50% to -30% (Maximum at Class 4)	**0**	6	2	2
+30% to -20% (Minimum at Class 5)	**11**	26	19	12
+20% to -15% (Minimum at Class 4)	**23**	39	29	24

5.2.2 Public Apartment Projects

1) Calculating Attribute Weights

In the same way as the military barracks projects, the attribute weights of public apartment projects cost estimating models were calculated by using four methods. In the cost estimating model of dual-optimization, the attribute weights were calculated by GA program 'Evolver' in a similar way to military barracks projects. Ten attribute weights and seven random variables were assigned as adjustable cells, and sum of distance was assigned as optimization cell. This process was performed for each fold, and a total of ten attribute weights sets were derived. The attribute weights and random variables calculation results of dual-optimization are summarized in Table 5-10.

Table 5-10 Calculation Results of Dual-Optimization (Public Apartment Projects)

Fold No.	Attribute weights									
	P1	P2	P3	P4	P5	P6	P7	Q1	Q2	Q3
1	0.1953	0.6860	0.0000	0.0000	0.0168	0.0000	0.0183	0.1019	0.0751	0.0004
2	0.2218	0.6541	0.0000	0.0000	0.0042	0.0000	0.0285	0.1023	0.0772	0.0000
3	0.0524	0.8212	0.0001	0.0000	0.0071	0.0000	0.0201	0.1108	0.0910	0.0001
4	0.1521	0.7131	0.0000	0.0000	0.0113	0.0000	0.0174	0.1119	0.0957	0.0096
5	0.2092	0.6630	0.0000	0.0000	0.0103	0.0000	0.0239	0.1046	0.0876	0.0031
6	0.0714	0.7930	0.0000	0.0000	0.0164	0.0000	0.0242	0.1126	0.0798	0.0000
7	0.2314	0.6581	0.0000	0.0000	0.0039	0.0000	0.0096	0.1076	0.0817	0.0009
8	0.4433	0.4355	0.0018	0.0000	0.0119	0.0000	0.0168	0.1006	0.0804	0.0050
9	0.0000	0.9124	0.0000	0.0000	0.0000	0.0000	0.0211	0.0771	0.0629	0.0000
10	0.2622	0.6120	0.0000	0.0000	0.0091	0.0000	0.0229	0.0989	0.0879	0.0079

Fold No.	Value of random variables for qualitative attributes						
	Q11	Q12	Q21	Q22	Q23	Q31	Q32
1	1.0000	0.0000	0.0000	0.3685	1.0000	0.0000	1.0000
2	1.0000	0.0000	0.0000	0.3575	1.0000	0.0000	0.0000
3	1.0000	0.0000	0.0000	0.3233	1.0000	0.0000	1.0000
4	1.0000	0.0000	0.0000	0.3136	1.0000	0.0000	1.0000
5	1.0000	0.0000	0.0000	0.3234	1.0000	0.0000	1.0000
6	1.0000	0.0000	0.0000	0.3507	1.0000	0.0000	0.0000
7	1.0000	0.0000	0.0000	0.3671	1.0000	0.0000	1.0000
8	1.0000	0.0000	0.0000	0.3498	1.0000	0.0000	1.0000
9	1.0000	0.0000	0.0000	0.4252	1.0000	0.0000	0.0000
10	1.0000	0.0000	0.0000	0.3204	1.0000	0.0000	1.0000

※ P1: Number of households, P2: Gross floor area, P3: Number of unit floor households, P4: Number of floors, P5: Number of elevators, P6: Number of household of unit floor per elevator, P7: Number of pilotis with household, Q1: Hallway type, Q2: Roof type, Q3: Building shape, Q11: Corridor, Q12, Stairway, Q21: Flat, Q22: Gable, Q23: Slope, Q31: Linear shape, Q32: L shape.

In the cost estimating model of feature counting, the weights of ten attributes were set to the same value of 0.1000 regardless of the fold. For the qualitative attributes, the binary method was applied to represent the distance between attribute values. The attribute weights of feature counting are summarized in Table 5-11.

Table 5-11 Attribute Weights of Feature Counting (Public Apartment Projects)

Attribute	P1	P2	P3	P4	P5	P6	P7	Q1	Q2	Q3
Weight	0.1000									

In order to calculate the attribute weights of regression analysis cost estimating model, regression analysis was performed for training set of each fold by setting the cost as a dependent variable and the seven quantitative variables as independent variables. The absolute value of the regression coefficient obtained through the analysis was used as the weight. The analysis was conducted by the program of MS Excel 2010 with confidence level of 95%. The attribute weights calculation results of regression analysis are summarized in Table 5-12.

Table 5-12 Calculation Results of Regression Analysis (Public Apartment Projects)

Fold No.	P1	P2	P3	P4	P5	P6	P7
1	0.0314	1.0847	0.0214	0.0927	0.0999	0.0808	0.0522
2	0.4406	0.6524	0.0022	0.0700	0.1263	0.0890	0.0602
3	0.4649	0.6386	0.0014	0.0768	0.1298	0.0943	0.0583
4	0.4418	0.6777	0.0082	0.0896	0.1280	0.0907	0.0551
5	0.6547	0.4984	0.0475	0.1112	0.1083	0.0915	0.0718
6	0.2643	0.8572	0.0150	0.0900	0.1260	0.0946	0.0649
7	0.5952	0.6059	0.0460	0.1131	0.1510	0.0999	0.0421
8	0.4369	0.6997	0.0319	0.1009	0.1031	0.0808	0.0501
9	0.5180	0.5848	0.0099	0.0716	0.1148	0.0874	0.0582
10	0.5789	0.4967	0.0167	0.0676	0.1332	0.0944	0.0673

In the cost estimating model of GA only quantitative attributes, the weights of seven quantitative attributes were calculated by GA program 'Evolver' as similar to the dual-optimization. The conditions for the GAs were a crossover rate of 0.05, a mutation rate of 0.1, an initial population of 50, and termination criterion of 5,000,000 trials reached. Seven quantitative attribute weights were assigned as adjustable cells, and sum of distance was assigned as optimization cell. The attribute weights calculation results of GA only quantitative attributes are summarized in Table 5-13.

Table 5-13 Calculation Results of GA only Quantitative Attributes (Public Apartment Projects)

Fold No.	P1	P2	P3	P4	P5	P6	P7
1	0.0000	0.9771	0.0000	0.0000	0.0000	0.0000	0.0336
2	0.0002	0.9778	0.0000	0.0000	0.0000	0.0000	0.0289
3	0.0000	0.9807	0.0000	0.0000	0.0000	0.0000	0.0278
4	0.0000	0.9797	0.0000	0.0000	0.0000	0.0000	0.0207
5	0.0000	0.9781	0.0000	0.0000	0.0000	0.0000	0.0289
6	0.0000	0.9770	0.0000	0.0000	0.0000	0.0000	0.0329
7	0.0000	0.9773	0.0000	0.0000	0.0000	0.0000	0.0325
8	0.0003	0.9801	0.0000	0.0000	0.0000	0.0000	0.0294
9	0.0001	0.9773	0.0000	0.0000	0.0000	0.0000	0.0325
10	0.0004	0.9712	0.0000	0.0000	0.0000	0.0000	0.0351

2) Results and Discussions

Based on the above attribute weights, 10-fold cross validation was performed. Forty CBR cost estimating models were developed and the construction costs of validation cases were estimated. Table 5-14 shows the comparison of MAERs by four weight calculation methods. In terms of estimation accuracy, the dual-optimization cost estimating model resulted in the lowest MAER at 7.01%, compared to feature counting at 9.33%,

regression analysis at 9.50%, and GA with only quantitative attributes at 8.01%. In the folds of 2, 3, 5, 6, 8 and 10, the dual-optimization method resulted in lowest MAER compared to the other methods. In the results of maximum MAER, the dual-optimization method resulted in the lowest value of 12.03% compared to feature counting at 16.05%, regression analysis at 18.03%, and GA with only quantitative attributes at 12.29%. Also, the dual-optimization method resulted in the lowest minimum MAER at 3.85%, compared to feature counting at 5.00%, regression analysis at 5.45%, and GA with only quantitative attributes at 4.81%.

Table 5-15 shows the number of cases beyond the expected accuracy range by four weights calculation methods. For the range from +100 to -50%, there was no case beyond the range. For the range from +50 to -30%, the dual-optimization and feature counting methods resulted that all cases were within the range although other methods resulted that there were one case out of range. In the results of other minimum ranges at Class 5 and 4, the dual-optimization method resulted in the lowest number of cases beyond the expected accuracy ranges compared to other methods. These results support that the dual-optimization method was more accurate than other methods.

Table 5-14 Comparison of MAERs by Weight Calculation Methods (Public Apartment Projects)

(MAER, %)

Fold No.	**Dual-optimization**	Feature Counting	Regression Analysis	GA only quantitative attributes
1	**8.23**	5.00	5.45	7.12
2	**3.85**	5.76	5.49	6.41
3	**6.09**	7.41	10.91	8.91
4	**6.35**	7.57	5.80	6.48
5	**4.69**	12.93	9.32	7.39
6	**12.03**	16.05	18.03	12.29
7	**9.63**	8.50	8.09	6.37
8	**5.10**	8.58	14.13	10.77
9	**7.64**	7.90	8.91	4.81
10	**6.52**	13.63	8.92	9.51
Average MAER	**7.01**	9.33	9.50	8.01
Max MAER	**12.03**	16.05	18.03	12.29
Min MAER	**3.85**	5.00	5.45	4.81

Table 5-15 Number of Cases beyond Expected Accuracy Range by Weight Calculation Methods (Public Apartment Projects)

Expected accuracy range	**Dual-optimization**	Feature Counting	Regression Analysis	GA only quantitative attributes
+100% to -50% (Maximum at Class 5)	**0**	0	0	0
+50% to -30% (Maximum at Class 4)	**0**	0	1	1
+30% to -20% (Minimum at Class 5)	**2**	6	7	7
+20% to -15% (Minimum at Class 4)	**10**	14	12	12

5.3 Validation 2: Retrieving Error Adaptation

Validation 2 was performed to verify the following hypothesis.

- Hypothesis 2: When the retrieving error adaptation method is applied to the GA-CBR cost estimating model, the estimation performance will be increased than when the method is not applied.

The validation was performed by comparing CBR outputs depending on whether the retrieving error adaptation was applied or not. For the five retrieved similar cases using the dual-optimization, the retrieving error adaptation was applied to adjust their solutions. Costs of validation cases were estimated to the mean of adjusted solutions of five retrieved cases. As similar to validation 1, 10-fold cross validation was performed and the effectiveness of the retrieving error adaptation was evaluated in terms of estimation accuracy by comparing the value of the MAER.

5.3.1 Military Barracks Projects

Table 5-16 shows the comparison of MAERs by with and without the retrieving error adaptation. When the retrieving error adaptation was applied to the GA-CBR cost estimating model, the average of MAERs decreased from 10.15% to 8.78%. In the all fords of test sets, the MAERs were reduced when the retrieving error adaptation was applied. The maximum and minimum MAERs were also reduced when the adaptation was applied from 13.53% to 12.26% and from 6.25% to 5.14%, respectively.

Table 5-17 shows the number of cases beyond the expected accuracy range by applying the retrieving error adaptation method. For the range from +100 to -50% and from +50 to -30%, there was no case beyond the range. For the range from +30 to -20% and from +20 to -15%, the numbers of cases beyond the expected range were decreased when the retrieving error adaptation was applied, from 11 to 8 and from 23 to 16, respectively. These results support that the estimation accuracy is improved when the retrieving error adaptation is applied.

Table 5-16 Comparison of MAERs by Applying Retrieving Error Adaptation (Military Barracks Projects)

(MAER, %)

Fold No.	Non-Applied	**Applied**	Variation
1	7.99	6.04	1.95
2	11.96	11.59	0.37
3	10.03	8.88	1.15
4	10.30	8.34	1.97
5	13.53	10.55	2.97
6	10.26	9.19	1.07
7	6.25	5.14	1.12
8	13.46	12.26	1.20
9	10.29	8.68	1.61
10	7.39	7.11	0.28
Average MAER	10.15	8.78	1.37
Max MAER	13.53	12.26	1.27
Min MAER	6.25	5.14	1.12

Table 5-17 Number of Cases beyond Expected Accuracy Range by Applying Retrieving Error Adaptation (Military Barracks Projects)

Expected accuracy range	Non-Applied	**Applied**
+100% to -50% (Maximum at Class 5)	0	**0**
+50% to -30% (Maximum at Class 4)	0	**0**
+30% to -20% (Minimum at Class 5)	11	**8**
+20% to -15% (Minimum at Class 4)	23	**16**

5.3.2 Public Apartment Projects

Table 5-18 shows the comparison of MAERs by with and without the retrieving error adaptation. As similar to the results of military barracks, when the retrieving error adaptation was applied to the GA-CBR cost estimating model, the average of MAERs decreased from 7.01% to 4.75%. In the all fords of test sets, the MAERs were reduced when the retrieving error adaptation was applied. The maximum and minimum MAERs were also reduced when the adaptation was applied from 12.03% to 9.24% and from 3.85% to 1.12%, respectively.

Table 5-19 shows the number of cases beyond the expected accuracy range by applying the retrieving error adaptation method. For the range from +100 to -50% and from +50 to -30%, there was no case beyond the range. For the range from +30 to -20% and from +20 to -15%, the numbers of cases beyond the expected range were decreased when the retrieving error adaptation was applied, from 2 to 1 and from 10 to 3, respectively. These results support that the estimation accuracy is improved when the retrieving error adaptation is applied.

Table 5-18 Comparison of MAERs by Applying Retrieving Error Adaptation (Public Apartment Projects)

(MAER, %)

Fold No.	Non-Applied	**Applied**	Variation
1	8.23	**3.83**	4.40
2	3.85	**1.12**	2.73
3	6.09	**3.36**	2.73
4	6.35	**5.94**	0.41
5	4.69	**4.03**	0.66
6	12.03	**7.65**	4.38
7	9.63	**9.24**	0.39
8	5.10	**4.76**	0.34
9	7.64	**3.05**	4.60
10	6.52	**4.58**	1.95
Average MAER	7.01	**4.75**	2.26
Max MAER	12.03	**9.24**	2.79
Min MAER	3.85	**1.12**	2.73

Table 5-19 Number of Cases beyond Expected Accuracy Range by Weight Calculation Methods (Public Apartment Projects)

Expected accuracy range	Non-Applied	**Applied**
+100% to -50% (Maximum at Class 5)	0	**0**
+50% to -30% (Maximum at Class 4)	0	**0**
+30% to -20% (Minimum at Class 5)	2	**1**
+20% to -15% (Minimum at Class 4)	10	**3**

5.4 Validation 3: Improved Weighted Mean Method

Validation 3 was performed to verify the following hypothesis.

- Hypothesis 3: When the improved weighted mean method is applied to the GA-CBR cost estimating model, the estimation accuracy will be superior to other existing methods.

The validation was performed by comparing CBR outputs of three multiple case adaptation methods: mean, weighted mean and improved weighted mean. Five similar cases were retrieved by using the dual-optimization; these were adapted by the retrieving error adaptation. Then, three multiple case adaptation methods were applied respectively to derive a solution from the obtained five solutions. Costs for validation cases were estimated, and the effectiveness of the improved weighted mean method was compared to other methods in terms of estimation accuracy by comparing the values of MAERs.

5.4.1 Military Barracks Projects

Table 5-20 shows the comparison of MAERs by multiple case adaptation methods. When the weighted mean method was applied to the GA-CBR cost estimating model, the average of MAERs increased 0.01%, from 8.78% to 8.79%. By comparison, when applying the improved weighted mean method, the MAERs were reduced in the folds of 2, 3, 4, 5, 6, and 9, and the MAER was also reduced 0.55%, from 8.78% to 8.23%. In conclusion, there was no significant change in the results when the mean and weighted mean methods were applied, but the estimation accuracy was increased relatively high when the improved weighted mean method was applied. Nevertheless, the maximum value of MAER was slightly increased and the numbers of cases beyond the expected accuracy range were remaining unchanged. By all accounts, when the weighted mean method was applied, there was little variation in the results, whereas the estimation accuracy was improved when the improved weighted mean method was applied.

Table 5-20 Comparison of MAERs by Multiple Case Adaptation Methods (Military Barracks Projects)

(MAER, %)

Fold No.	Mean	Weighted mean	**Improved weighted mean**
1	6.04	6.07	**6.77**
2	11.59	11.62	**8.53**
3	8.88	8.89	**8.80**
4	8.34	8.38	**7.64**
5	10.55	10.55	**10.14**
6	9.19	9.19	**6.78**
7	5.14	5.12	**5.14**
8	12.26	12.26	**12.54**
9	8.68	8.69	**8.37**
10	7.11	7.11	**7.54**
Average MAER	8.78	8.79	**8.23**
Max MAER	12.26	12.26	**12.54**
Min MAER	5.14	5.12	**5.14**

Table 5-21 Number of Cases beyond Expected Accuracy Range by Multiple Case Adaptation Methods (Military Barracks Projects)

Expected accuracy range	Mean	Weighted mean	**Improved weighted mean**
+100% to -50% (Maximum at Class 5)	**0**	**0**	**0**
+50% to -30% (Maximum at Class 4)	**0**	**0**	**0**
+30% to -20% (Minimum at Class 5)	**8**	**8**	**8**
+20% to -15% (Minimum at Class 4)	**16**	**16**	**16**

5.4.2 Public Apartment Projects

Table 5-22 shows the comparison of MAERs by multiple case adaptation methods. When the weighted mean method was applied to the GA-CBR cost estimating model, the average of MAERs decreased only 0.01%, from 4.75% to 4.74%. By comparison, when applying the improved weighted mean method, the MAERs were reduced in the folds of 1, 2, 6, 9 and 10, and the MAER was also reduced 0.26%, from 4.75% to 4.49%. In conclusion, there was no significant change in the results when the mean and weighted mean methods were applied, but the estimation accuracy was increased relatively high when the improved weighted mean method was applied. Nevertheless, the maximum value of MAER was slightly increased and the numbers of cases beyond the expected accuracy range were remaining unchanged. By all accounts, when the weighted mean method was applied, there was little variation in the results, whereas the estimation accuracy was improved when the improved weighted mean method was applied. As compared with the result of the military barracks, the variation ranges of the results of the public apartments were relatively small because the case base of public apartment was composed of similar cases to each other.

Table 5-22 Comparison of MAERs by Multiple Case Adaptation Methods (Public Apartment Projects)

(MAER, %)

Fold No.	Mean	Weighted mean	**Improved weighted mean**
1	6.04	6.07	**6.77**
2	11.59	11.62	**8.53**
3	8.88	8.89	**8.80**
4	8.34	8.38	**7.64**
5	10.55	10.55	**10.14**
6	9.19	9.19	**6.78**
7	5.14	5.12	**5.14**
8	12.26	12.26	**12.54**
9	8.68	8.69	**8.37**
10	7.11	7.11	**7.54**
Average MAER	8.78	8.79	**8.23**
Max MAER	12.26	12.26	**12.54**
Min MAER	5.14	5.12	**5.14**

Table 5-23 Number of Cases beyond Expected Accuracy Range by Multiple Case Adaptation Methods (Public Apartment Projects)

Expected accuracy range	Mean	Weighted mean	**Improved weighted mean**
+100% to -50% (Maximum at Class 5)	**0**	**0**	**0**
+50% to -30% (Maximum at Class 4)	**0**	**0**	**0**
+30% to -20% (Minimum at Class 5)	**1**	**1**	**1**
+20% to -15% (Minimum at Class 4)	**3**	**3**	**3**

5.5 Summary

In order to validate three methods proposed in chapter 3, three kinds of validations were conducted to estimate the construction cost of military barracks and public apartment projects by the 10-fold cross validation method. By comparing the average of MAERs, when the proposed methods were applied, it was analyzed whether the estimation accuracy improved compared with the existing methods. In the validation 1, the performance of the dual-optimization method was compared to other weight calculation methods: feature counting, regression analysis, and GA method using only quantitative attributes. As a result of the validation, although there were slightly differences for individual test set, the estimation accuracy of the dual-optimization was better than that of other methods. In the validation 2, the effectiveness of the retrieving error adaptation method was evaluated by comparing CBR outputs depending on whether the retrieving error adaptation was applied or not. The estimation accuracy were improved when retrieving error adaptation was applied into two GA-CBR cost estimating model. In validation 3, the improved weighted mean method was compared to other methods by comparing CBR outputs of three multiple case adaptation methods: mean, weighted mean and improved weighted mean. When the weighted mean method was applied, there was little

variation in the results, whereas the estimation accuracy was increased when the improved weighted mean method was applied.

Table 5-24 is summary of the three validation results. Compared with the conventional GA method, when considering the qualitative attributes by the dual-optimization method, the average of MAERs in both cost estimating model were reduced from 11.07 to 10.15% and 8.01 to 7.01%. Also, when adapting the differences between a problem and retrieved similar cases by using the retrieving error adaptation method, the average of MAERs decreased from 10.15 to 8.78% and 7.01 to 4.75%. For multiple case adaptations, the estimation accuracy was further enhanced by using the improved weighted mean adaptation method, from 8.78 to 8.23% and from 4.75 to 4.49%. Finally, when three proposed methods were applied together, the estimation accuracy of two GA-CBR cost estimating models increased by 2.84% and 3.52% compared to the conventional GA method. Additionally, for the range expected at the highest level in the early stage cost estimation (+20 to -15%), the number of cases beyond the range were reduced from 24 to 16 and from 12 to 3. These results means the proposed methods have improved accuracy in comparison with existing methods and can be fully utilized for the cost estimation

Table 5-24 Summary of Validation Results

Applied Method	Military barrack projects		Public apartment projects	
	Average MAER (%)	No. of cases beyond range	Average MAER (%)	No. of cases beyond range
GA only quantitative attributes	**11.07**	**24**	**8.01**	**12**
Dual-optimization	**10.15**	**23**	**7.01**	**10**
Dual-optimization + Retrieving error adaptation	**8.78**	**16**	**4.75**	**3**
Dual-optimization + Retrieving error adaptation + Improved weighted mean	**8.23**	**16**	**4.49**	**3**

Chapter 6. Conclusions

6.1 Summary of Research

Early stage cost estimation has a significant effect on the success of project, whereas the information is restricted. Therefore, various methods for improving the estimation accuracy have been studied and applied. Among them, case-based reasoning is one of the most commonly utilized methods because it is able to produce persuasive solutions by applying a similar reasoning process to that of human problem solving when limited information is provided.

The accuracy of the case-based reasoning cost estimating model has been affected by retrieving and adaptation. In order to improve the retrieving performance, various methods have been suggested for assigning attribute weights in case-based reasoning. However, previous methods of assigning attribute weights are limited when used to calculate the attribute weights of qualitative variables. Hence, this limits the types of variables that may be used in case-based reasoning. To address this problem, this research proposed the new method, termed the dual-optimization method. Based on genetic algorithm, it can assign not only the attribute weights of

both types of variables, but also the quantified attribute values of the qualitative variables. Additionally, this research suggested two adaptation methods for the GA-CBR cost estimating model. The retrieving error adaptation is the method to calculate the differences and to adjust the solutions of retrieved cases. By reflecting estimation error caused by differences between target and retrieved cases, a retrieved solution is adjusted to be more appropriate. Furthermore, the improved weighted mean method was suggested to alteration method for multiple cases adaptation. By assigning weights according to the similarity distribution of retrieved cases, the improved weighted mean method can increase the influence of more similar cases.

To validate the proposed methods, three kinds of validations were conducted to estimate the construction cost of military barracks and public apartment projects. The results of validation 1 indicate that the dual-optimization method is improved in terms of accuracy and stability compared to previous methods. In validation 2 and 3, the estimation accuracy is increased when the proposed adaptation methods are used to the GA-CBR cost estimating model. These validation results support that the proposed methods can be utilized for construction cost estimation to make better decision.

6.2 Research Contributions

This research has significant in that it suggests three new methods to improve the disadvantages of existing methods. The expected contributions of this research can be summarized as follows.

1) The dual-optimization method can be distinguished from previous approaches in that calculating weights and quantifying qualitative attributes are performed concurrently. It is possible to expand the range of applicable attribute for CBR, whereas this was limited before.

2) The retrieving error adaptation method can modify the solutions without additional analysis or database. By adding the error to the solution of each case, the retrieved cases are adjusted to fit to the problem description.

3) The improved mean method can increase the influence of more similar cases. By increasing the influence of more similar cases, the overall solution will be more accurate.

Consequently, the new GA-CBR cost estimating model can make more accurate cost estimation compared to other methods. It is expected that the proposed cost estimating model and methods will support stakeholders of construction project to make better decision at early stage of project. Moreover, the dual-optimization is a general-purpose method; it is expected to be more readily applied to a problem of other fields. This research used qualitative variables on a nominal scale as an example, but the approach is also applicable to qualitative variables on ordinal scales such as finishing grades and intelligence levels.

6.3 Limitations and Further Studies

In the dual-optimization, despite of its excellent performance, the more qualitative variables and values are utilized, the longer the length of the chromosome to be optimized and the longer the calculation duration. This research hopes that this problem will be solved by advance of computer science and improvement of its algorithm. Because this research was conducted by using Korean military barracks and public apartment projects, additional validation will be required when the proposed approach is applied to other targets. Since this research is focused on the retrieve and reuse process, enhancing of the usability of the GA-CBR cost estimating model through research on the revise and retain process will be necessary in further studies.

Bibliography

Aamodt, A., & Plaza, E. (1994). "Case-based reasoning: Foundational issues, methodological variations, and system approaches." AI communications, 7(1), 39-59.

Ahn, J. (2016). "Front-end cost estimation by selective case-based reasoning for building construction projects." Ph. D. dissertation, Seoul National Univ., Seoul, South Korea.

Ahn, J., Ji, S. H., Park, M., Lee, H. S., Kim, S., & Suh, S. W. (2014). "The attribute impact concept: Applications in case-based reasoning and parametric cost estimation." *Automation in Construction*, *43*, 195-203.

Ahuja, H. N., Dozzi, S. P., & Abourizk, S. M. (1994). Project management: techniques in planning and controlling construction projects. John Wiley & Sons.

An, S. H., Kim, G. H., & Kang, K. I. (2007). "A case-based reasoning cost estimating model using experience by analytic hierarchy process." *Building and Environment*, *42*(7), 2573-2579.

Arafa, M., & Alqedra, M. (2011). "Early stage cost estimation of buildings construction projects using artificial neural networks." *Journal of Artificial Intelligence*, *4*(1), 63-75.

Arditi, D., & Tokdemir, O. B. (1999). "Comparison of case-based reasoning and artificial neural networks." *Journal of computing in civil engineering*, *13*(3), 162-169.

Baker, J. E. (1985). Adaptive selection methods for genetic algorithms. *International Conference on Genetic Algorithms and their applications,* 101-111.

Barletta, R. (1991). "An introduction to case-based reasoning." *AI expert*, *6*(8), 42-49.

Barrie, D. S., & Paulson Jr, B. C. (2000). Professional construction management: including CM, design-construct, and general contracting. McGraw-Hill. 3rd edition.

Begum, S., Ahmed, M. U., Funk, P., Xiong, N., & Von Schéele, B. (2009). "A case-based decision support system for individual stress diagnosis using fuzzy similarity matching." *Computational Intelligence*, *25*(3), 180-195.

Brown, C. E., & Gupta, U. G. (1994). "Applying Case-Based Reasoning to the Accounting Domain." *Intelligent Systems in Accounting, Finance and Management*, *3*(3), 205-221.

Burkhard, H. D. (2001). "Similarity and distance in case based reasoning." *Fundamenta Informaticae*, *47*(3-4), 201-215.

Christensen, P., & Dysert, L. R. (2005). Cost estimate classification system–as applied in engineering, procurement, and construction for the process industries. *Morgantown, WV: AACE International.*

Chua, D. K. H., Li, D. Z., & Chan, W. T. (2001). "Case-based reasoning approach in bid decision making." *Journal of construction engineering and management*, *127*(1), 35-45.

Chun, S. H., & Park, Y. J. (2006). "A new hybrid data mining technique using a regression case based reasoning: Application to financial forecasting." *Expert Systems with Applications*, *31*(2), 329-336.

De Mantaras, R. L., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Aamodt, A., & Watson, I. (2005). "Retrieval, reuse, revision and retention in case-based reasoning." *The Knowledge Engineering Review*, *20*(3), 215-240.

Dikmen, I., Birgonul, M. T., & Gur, A. K. (2007). "A case-based decision support tool for bid mark-up estimation of international construction projects." *Automation in Construction*, *17*(1), 30-44.

Doğan, S. Z., Arditi, D., & Günaydın, 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." *The international journal of advanced manufacturing technology*, *15*(12), 895-906.

Elhag, T. M. S., & Boussabaine, A. H. (1998). "An artificial neural system for cost estimation of construction projects." *14th ARCOM Annual Conference,* 219-226.

Fushiki, T. (2011). "Estimation of prediction error by using K-fold cross-validation." *Statistics and Computing*, *21*(2), 137-146.

Garza, A. D. G. S., & Maher, M. L. (1999). "An evolutionary approach to case adaptation." *International Conference on Case-Based Reasoning 1999,* 162-173.

Gen, M., & Cheng, R. (2000). *Genetic algorithms and engineering optimization*. John Wiley & Sons.

Goh, Y. M., & Chua, D. K. H. (2009). "Case-based reasoning approach to construction safety hazard identification: adaptation and utilization." *Journal of Construction Engineering and Management*, *136*(2), 170-178.

Golberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. *Addion wesley*, *1989*, 102.

Günaydın, H. M., & Doğan, S. Z. (2004). "A neural network approach for early cost estimation of structural systems of buildings." *International Journal of Project Management*, *22*(7), 595-602.

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.

Holland, J. H. (1975). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press.

Hong, T., Hyun, C., & Moon, H. (2011). “CBR-based cost prediction model-II of the design phase for multi-family housing projects.” *Expert Systems with Applications*, *38*(3), 2797-2808.

Hu, J., Qi, J., & Peng, Y. (2015). “New CBR adaptation method combining with problem–solution relational analysis for mechanical design.” *Computers in Industry*, *66*, 41-51.

Ji, C., Hong, T., & Hyun, C. (2010). “CBR revision model for improving cost prediction accuracy in multifamily housing projects.” *Journal of Management in Engineering*, *26*(4), 229-236.

Ji, S. H., Park, M., Lee, H. S., Ahn, J., Kim, N., & Son, B. (2010). “Military facility cost estimation system using case-based reasoning in Korea.” *Journal of Computing in Civil Engineering*, *25*(3), 218-231.

Ji, S. H. (2011). “Improvement of retrieval and adaptation methods in case-based reasoning for construction cost estimation.” Ph. D. dissertation, Seoul National Univ., Seoul, South Korea.

Ji, S. H., Park, M., & Lee, H. S. (2011). “Cost estimation model for building projects using case-based reasoning.” *Canadian Journal of Civil Engineering*, *38*(5), 570-581.

Kim, G., & Kang, K. (2004). "A study on predicting construction cost of apartment housing projects based on case based reasoning technique at the early project stage." *Journal of Architectural Institute of Korea*, *20*(5), 83-92.

Kim, K. J., & Kim, K. (2010). "Preliminary cost estimation model using case-based reasoning and genetic algorithms." *Journal of Computing in Civil Engineering*, *24*(6), 499-505.

Kolodner, J. (1993). *Case-based reasoning*, Morgan Kaufmann Publishers Inc., New York, USA.

Koo, C., Hong, T., & Hyun, C. (2011). "The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach." *Expert Systems with Applications*, *38*(7), 8597-8606.

Korea Institute of Construction Technology, Construction Cost Index, 2016.

Leake, D. B. (1996). Case-Based Reasoning: Experiences, lessons and future directions. MIT press.

Liao, Z., Mao, X., Hannam, P. M., & Zhao, T. (2012). "Adaptation methodology of CBR for environmental emergency preparedness system

based on an Improved Genetic Algorithm." *Expert Systems with Applications*, *39*(8), 7029-7040.

Lotfy, E. A., & Mohamed, A. S. (2002). "Applying neural networks in case-based reasoning adaptation for cost assessment of steel buildings." *International journal of computers and applications*, *24*(1), 28-38.

Luu, D. T., Ng, S. T., & Chen, S. E. (2005). "Formulating procurement selection criteria through case-based reasoning approach." *Journal of computing in civil engineering*, *19*(3), 269-276.

McLachlan, G., Do, K. A., & Ambroise, C. (2005). *Analyzing microarray gene expression data*. John Wiley & Sons.

Montazemi, A. R., & Gupta, K. M. (1997). "A framework for retrieval in case-based reasoning systems." *Annals of operations research*, *72*, 51-73.

Park, C. S., & Han, I. (2002). "A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction." *Expert Systems with Applications*, *23*(3), 255-264.

Patterson, D. W., Rooney, N., & Galushka, M. (2003). "Efficient Retrieval for Case-Based Reasoning." In *FLAIRS Conference,* 144-149.

Paulson Jr, B. C. (1976). Designing to reduce construction costs. *Journal of the construction division*, *102*(C04).

Policastro, C. A., Carvalho, A. C., & Delbem, A. C. (2003). "Hybrid approaches for case retrieval and adaptation." *Annual Conference on Artificial Intelligence*, 297-311.

Policastro, C. A., Carvalho, A. C., & Delbem, A. C. (2008). "A hybrid case adaptation approach for case-based reasoning." *Applied Intelligence*, *28*(2), 101-119.

Qi, J., Hu, J., & Peng, Y. (2012). "A new adaptation method based on adaptability under k-nearest neighbors for case adaptation in case-based design." *Expert Systems with Applications*, *39*(7), 6485-6502.

Richter, M. M., & Weber, R. O. (2013). "Complex Similarity Topics." *Case-Based Reasoning,* 149-165.

Riesbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*. Lawrence Earlbaum, Hillsdale, NJ.

Ryu, H. G., Lee, H. S., & Park, M. (2007). “Construction planning method using case-based reasoning (CONPLA-CBR).” *Journal of Computing in Civil Engineering*, *21*(6), 410-422.

Seni, G., & Elder, J. F. (2010). “Ensemble methods in data mining: improving accuracy through combining predictions.” *Synthesis Lectures on Data Mining and Knowledge Discovery*, *2*(1), 1-126.

Sharifi, M., Naghibzadeh, M., & Rouhani, M. (2013). “Adaptive case-based reasoning using support vector regression.” *Advance Computing Conference (IACC), 2013 IEEE 3rd International,* 1006-1010.

Shiu, S. C., & Pal, S. K. (2004). “Case-based reasoning: concepts, features and soft computing.” *Applied Intelligence*, *21*(3), 233-238.

Trost, S. M., & Oberlender, G. D. (2003). “Predicting accuracy of early cost estimates using factor analysis and multivariate regression.” *Journal of Construction Engineering and Management*, *129*(2), 198-204.

Turban, E., & Frenzel, L. E. (1992). *Expert systems and applied artificial intelligence*. Prentice Hall Professional Technical Reference.

Watson, I., & Marir, F. (1994). “Case-based reasoning: A review.” *Knowledge Engineering Review*, *9*(4), 327-354.

Watson, I. (1999). “Case-based reasoning is a methodology not a technology.” *Knowledge-based systems*, *12*(5), 303-308.

Yau, N. J., & Yang, J. B. (1998). “Case-based reasoning in construction management.” *Computer-Aided Civil and Infrastructure Engineering*, *13*(2), 143-150.

Appendix

Appendix 1. Grossary of Acronyms

Acronyms	Fullname
AACE	Association for the Advancement of Cost Engineering
AER	Absolute Error Ratio
AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Network
CBR	Case-based Reasoning
CI	Cost Index
FC	Feature Counting
GA	Genetic Algorithm
KICT	Korea Institute of Civil Engineering and Building Technology
k-fold CV	k-fold Cross Validation
k-NN	k-Nearest Neighbor
MAER	Mean of Absolute Error Ratio
RA	Regression Analysis
RBR	Rule-based Reasoning

Appendix 2. Case Base of Military Barrack Projects

Case	P1	P2	P3	P4	Q1	Q2	Cost (2008)
Case 1	72	2	1,939.0	950.0	navy	RC	2,232,454,819
Case 2	60	1	1,007.5	1,007.5	army	mixed	982,905,211
Case 3	45	1	779.7	779.7	army	mixed	778,416,569
Case 4	74	1	1,093.8	1,093.8	army	mixed	1,059,117,935
Case 5	40	1	570.9	570.9	army	mixed	612,837,620
Case 6	81	2	2,660.0	1,339.0	army	RC	2,154,812,322
Case 7	40	1	596.0	595.2	army	mixed	586,822,936
Case 8	84	1	978.0	978.0	army	mixed	962,676,600
Case 9	128	3	1,593.2	616.1	air force	RC	1,407,956,647
Case 10	36	2	603.8	341.0	army	mixed	738,387,238
Case 11	28	1	462.9	462.9	army	mixed	393,700,403
Case 12	80	2	1,061.9	538.7	air force	RC	1,005,144,727
Case 13	80	2	1,061.9	538.7	air force	RC	978,190,737
Case 14	104	2	1,299.2	657.4	air force	RC	958,866,410
Case 15	380	3	7,457.2	2,515.0	navy	RC	6,522,553,727

Case 16	192	3	6,272.8	2,927.7	navy	RC	6,728,168,070
Case 17	42	1	706.9	740.2	air force	RC	853,459,099
Case 18	252	3	2,694.3	1,416.0	air force	RC	3,407,455,314
Case 19	71	1	1,144.6	1,145.0	army	steel	1,270,249,147
Case 20	47	1	717.8	717.8	army	steel	841,703,538
Case 21	68	2	4,988.9	2,531.0	navy	RC	5,831,180,159
Case 22	28	1	431.7	408.3	air force	RC	632,894,851
Case 23	14	1	709.9	671.3	air force	RC	799,490,299
Case 24	81	2	1,874.7	953.0	army	RC	1,836,004,834
Case 25	72	1	1,060.4	547.0	air force	RC	990,905,708
Case 26	145	3	2,003.2	755.1	air force	RC	1,887,601,103
Case 27	105	2	1,298.3	677.1	air force	RC	1,403,268,646
Case 28	105	4	1,521.6	465.9	air force	RC	1,648,713,998
Case 29	21	1	293.8	295.3	air force	RC	414,071,413
Case 30	40	2	1,666.3	786.9	army	RC	1,601,554,394
Case 31	20	1	393.0	393.0	army	steel	514,260,789
Case 32	122	3	2,559.0	858.0	army	RC	2,228,077,197
Case 33	8	1	82.6	82.6	army	steel	99,236,827
Case 34	189	3	6,097.3	2,123.4	army	RC	4,772,718,059

Case 35	142	3	2,460.0	822.0	army	RC	2,347,045,476
Case 36	14	2	890.5	486.0	air force	RC	1,205,039,379
Case 37	289	4	5,550.2	1,382.2	army	RC	4,538,263,785
Case 38	24	3	801.6	322.8	army	RC	870,151,956
Case 39	392	3	6,917.8	1,976.0	army	RC	5,366,632,215
Case 40	106	2	1,873.0	962.2	army	RC	1,912,050,360
Case 41	54	3	959.0	319.7	army	RC	1,010,128,234
Case 42	105	3	1,439.5	493.0	air force	RC	1,450,706,367
Case 43	311	3	5,249.2	2,049.0	air force	RC	4,321,460,296
Case 44	98	2	1,906.1	1,059.8	army	RC	2,180,510,614
Case 45	72	3	1,932.5	942.2	army	mixed	1,672,863,927
Case 46	35	2	579.1	289.5	army	RC	560,890,799
Case 47	12	1	240.0	240.0	army	mixed	300,398,549
Case 48	102	3	2,635.4	1,056.9	army	mixed	2,453,940,556
Case 49	28	2	1,644.2	899.2	army	RC	1,476,664,259
Case 50	37	2	842.3	453.9	army	steel	699,124,815
Case 51	96	2	1,906.1	1,059.8	army	RC	1,989,505,600
Case 52	224	3	4,906.3	1,709.8	army	RC	3,914,974,341
Case 53	56	2	864.9	440.5	air force	RC	865,879,948

Case 54	60	2	860.6	439.8	air force	RC	876,372,953
Case 55	466	3	9,160.1	3,377.9	army	RC	7,490,701,837
Case 56	304	4	6,535.6	2,094.1	army	RC	5,747,221,559
Case 57	38	2	725.6	356.2	army	steel	664,267,799
Case 58	41	1	599.2	599.2	army	steel	618,107,622
Case 59	260	3	3,462.3	1,325.5	army	RC	2,885,350,217
Case 60	189	3	5,335.1	2,021.6	army	RC	4,373,543,326
Case 61	22	2	381.8	193.8	army	mixed	421,302,945
Case 62	50	1	714.2	714.2	army	steel	773,864,444
Case 63	40	2	828.0	414.0	army	steel	928,306,982
Case 64	40	1	559.4	559.4	army	steel	616,522,015
Case 65	74	2	1,143.6	573.6	army	steel	1,146,976,197
Case 66	80	1	978.0	978.0	army	steel	1,017,099,099
Case 67	45	1	605.8	605.8	army	steel	635,314,759
Case 68	43	1	745.0	745.0	army	steel	787,863,581
Case 69	90	2	1,482.6	794.4	army	RC	910,518,466
Case 70	42	2	760.0	388.0	army	steel	873,036,776
Case 71	35	2	629.8	452.2	army	steel	878,652,957
Case 72	200	3	3,308.3	1,064.2	army	RC	2,021,049,246

Case 73	200	3	3,322.9	1,235.0	army	RC	2,296,422,647
Case 74	203	2	3,201.0	1,633.5	army	RC	2,733,754,560
Case 75	108	3	2,476.0	946.7	army	RC	2,443,040,652
Case 76	80	2	1,977.2	988.6	army	RC	1,880,948,706
Case 77	12	2	550.2	262.8	army	RC	626,516,217
Case 78	92	3	3,408.3	1,116.8	army	RC	2,395,170,597
Case 79	452	3	2,704.1	831.8	army	RC	2,708,635,338
Case 80	220	3	3,657.6	1,258.2	army	RC	2,164,848,129
Case 81	200	3	3,687.7	1,245.0	army	RC	3,268,584,148
Case 82	327	4	8,268.1	1,884.2	army	RC	7,138,933,135
Case 83	239	3	5,303.8	1,870.9	army	RC	4,153,541,328
Case 84	289	3	5,550.2	1,382.2	army	RC	4,695,451,147
Case 85	40	2	618.1	314.5	army	steel	759,767,923
Case 86	70	2	1,450.3	808.0	army	RC	1,475,507,606
Case 87	59	2	1,066.0	569.4	army	mixed	895,946,812
Case 88	260	3	5,434.5	2,213.7	army	mixed	4,460,670,574
Case 89	190	3	3,133.7	1,168.2	navy	RC	2,851,931,953
Case 90	105	3	1,516.2	531.5	air force	RC	1,707,973,500
Case 91	42	2	622.3	318.4	air force	RC	903,627,735

Case 92	26	1	647.1	544.2	army	RC	597,404,120
Case 93	98	2	1,306.0	668.6	air force	RC	1,471,210,150
Case 94	168	3	2,724.6	941.8	air force	RC	2,717,242,448
Case 95	40	1	609.1	609.1	army	mixed	585,815,455
Case 96	36	1	498.8	498.8	army	steel	574,470,703
Case 97	80	2	1,061.9	538.7	air force	RC	956,429,203
Case 98	90	2	2,161.8	1,097.4	navy	RC	2,072,248,449
Case 99	48	3	1,222.8	419.8	air force	RC	1,229,501,966
Case 100	216	3	2,740.1	903.8	army	RC	2,578,185,135
Case 101	105	2	1,298.3	677.1	air force	RC	1,363,288,227
Case 102	43	1	860.0	770.0	army	steel	1,021,887,293
Case 103	240	3	3,087.7	1,079.3	air force	RC	2,899,324,864
Case 104	50	2	1,031.6	444.0	army	RC	1,053,957,880
Case 105	30	2	628.4	322.2	army	steel	747,314,167
Case 106	40	1	606.9	607.2	army	mixed	575,027,275
Case 107	56	2	824.4	431.9	air force	RC	1,041,030,865
Case 108	266	3	5,494.8	1,771.6	army	RC	4,294,238,795
Case 109	117	2	1,874.7	921.7	army	RC	1,855,033,273
Case 110	47	2	760.7	382.3	army	steel	840,565,210

Case 111	44	2	785.0	396.8	army	steel	834,860,637
Case 112	55	2	907.2	576.6	army	steel	976,255,332
Case 113	80	1	978.0	978.0	army	steel	1,004,394,438
Case 114	56	2	747.0	378.6	air force	RC	959,287,848

※ P1: Capacity, P2: Number of floors, P3: Gross floor area, P4: Building area, Q1: Military type, Q2: Structure type

Appendix 3. Case Base of Public Apartment Projects

Case	P1	P2	P3	P4	P5	P6	P7	Q1	Q2	Q3	Cost(2008)
Case 1	46	5,065.7	4	12	1	4	4	Corridor	Flat	L shape	2,670,302,790
Case 2	24	2,638.6	2	12	1	2	0	Stairway	Slope	Linear shape	1,478,020,520
Case 3	48	5,277.1	4	12	2	2	0	Stairway	Slope	Linear shape	2,647,914,525
Case 4	48	5,277.1	4	12	2	2	0	Stairway	Slope	Linear shape	2,629,430,221
Case 5	40	4,448.1	4	11	1	4	4	Corridor	Flat	L shape	2,398,184,147
Case 6	44	4,840.0	4	11	2	2	0	Stairway	Slope	Linear shape	2,393,077,377
Case 7	50	5,500.4	4	13	1	4	2	Corridor	Flat	L shape	2,668,546,560
Case 8	56	6,167.5	4	15	1	4	4	Corridor	Flat	L shape	3,091,669,805
Case 9	54	5,932.8	4	14	2	2	0	Stairway	Slope	Linear shape	2,905,716,406
Case 10	54	5,932.8	4	14	2	2	0	Stairway	Slope	Linear shape	2,903,619,607
Case 11	44	4,907.9	4	12	2	2	4	Stairway	Slope	Linear shape	2,576,905,089
Case 12	54	5,950.4	4	15	1	4	2	Corridor	Flat	L shape	3,064,384,232
Case 13	52	5,750.1	4	15	1	4	4	Corridor	Flat	L shape	3,059,002,015
Case 14	52	5,750.1	4	15	1	4	4	Corridor	Flat	L shape	3,060,696,280
Case 15	20	2,200.6	2	10	1	2	0	Stairway	Slope	Linear shape	1,187,349,129

Case 16	28	3,108.1	2	15	1	2	2	Stairway	Slope	Linear shape	1,652,419,018
Case 17	56	6,189.1	4	15	1	4	4	Corridor	Flat	L shape	3,028,052,374
Case 18	56	6,189.1	4	15	1	4	4	Corridor	Flat	L shape	3,050,837,246
Case 19	52	5,753.2	4	14	1	4	4	Corridor	Flat	L shape	3,005,099,417
Case 20	28	3,103.3	2	15	1	2	2	Stairway	Slope	Linear shape	1,628,612,706
Case 21	50	5,551.9	4	14	2	2	4	Stairway	Slope	Linear shape	2,832,633,540
Case 22	44	4,897.6	4	12	2	2	4	Stairway	Slope	Linear shape	2,619,514,629
Case 23	22	2,496.5	4	8	2	2	4	Stairway	Slope	Linear shape	1,569,350,543
Case 24	56	6,163.6	4	15	1	4	4	Corridor	Flat	L shape	3,288,064,443
Case 25	50	5,501.4	4	13	1	4	2	Corridor	Flat	L shape	2,947,000,635
Case 26	50	5,501.4	4	13	1	4	2	Corridor	Flat	L shape	2,898,704,154
Case 27	24	2,655.5	2	13	1	2	2	Stairway	Flat	Linear shape	1,407,473,903
Case 28	16	1,754.1	2	8	1	2	0	Stairway	Slope	Linear shape	792,836,552
Case 29	22	2,541.7	2	14	1	2	0	Stairway	Flat	Linear shape	1,148,868,556
Case 30	20	2,190.0	2	10	1	2	0	Stairway	Flat	Linear shape	1,005,030,335
Case 31	16	1,809.2	2	9	1	2	2	Stairway	Slope	L shape	920,592,438
Case 32	18	2,021.1	2	9	1	2	0	Stairway	Slope	Linear shape	977,962,918
Case 33	22	2,410.4	2	11	1	2	0	Stairway	Slope	Linear shape	1,166,349,647
Case 34	20	2,244.7	2	10	1	2	0	Stairway	Slope	Linear shape	1,061,027,687

Case 35	24	2,627.2	2	12	1	2	0	Stairway	Slope	Linear shape	1,241,821,137
Case 36	26	2,879.9	2	14	1	2	2	Stairway	Flat	Linear shape	1,295,871,949
Case 37	5	555.1	1	7	0.5	2	2	Stairway	Gable	Linear shape	343,286,007
Case 38	12	1,293.0	2	6	1	2	0	Stairway	Gable	L shape	881,914,132
Case 39	7	735.2	1	7	0.5	2	0	Stairway	Gable	Linear shape	446,315,141
Case 40	14	1,507.6	2	7	1	2	0	Stairway	Gable	L shape	1,006,531,964
Case 41	7	735.2	1	7	0.5	2	0	Stairway	Gable	Linear shape	445,348,522
Case 42	13	1,415.8	2	7	1	2	1	Stairway	Gable	L shape	970,004,637
Case 43	13	1,422.8	2	7	1	2	1	Stairway	Gable	L shape	942,560,605
Case 44	6	646.3	1	6	0.5	2	1	Stairway	Gable	Linear shape	357,510,753
Case 45	7	735.2	1	7	0.5	2	0	Stairway	Gable	Linear shape	423,322,113
Case 46	7	735.2	1	7	0.5	2	0	Stairway	Gable	Linear shape	404,340,643
Case 47	10	1,078.4	2	5	1	2	0	Stairway	Gable	L shape	672,515,594
Case 48	14	1,507.6	2	7	1	2	0	Corridor	Gable	L shape	916,809,152
Case 49	23	2,493.6	2	12	1	2	1	Stairway	Flat	Linear shape	1,028,807,009
Case 50	24	2,529.5	2	12	1	2	0	Stairway	Flat	Linear shape	1,164,726,244
Case 51	11	1,174.8	2	6	1	2	1	Stairway	Flat	Linear shape	540,958,492
Case 52	22	2,325.1	2	11	1	2	0	Stairway	Flat	Linear shape	1,113,068,832
Case 53	22	2,365.1	2	11	0.5	4	0	Stairway	Flat	Linear shape	1,042,935,423

Case 54	22	2,365.1	2	11	0.5	4	0	Stairway	Flat	Linear shape	1,042,373,304
Case 55	20	2,150.1	2	10	0.5	4	0	Stairway	Flat	Linear shape	946,513,442
Case 56	20	2,150.1	2	10	0.5	4	0	Stairway	Flat	Linear shape	972,380,452
Case 57	28	3,010.9	4	7	2	2	0	Stairway	Flat	Linear shape	1,262,732,352
Case 58	26	3,172.2	4	7	2	2	2	Stairway	Flat	L shape	1,245,140,948
Case 59	23	2,466.3	4	6	2	2	1	Stairway	Flat	Linear shape	1,116,215,830
Case 60	24	2,556.7	4	6	2	2	0	Stairway	Flat	Linear shape	1,106,457,137
Case 61	24	2,556.7	4	6	2	2	0	Stairway	Flat	Linear shape	1,109,508,468
Case 62	15	1,623.1	4	4	2	2	1	Stairway	Flat	Linear shape	742,757,532
Case 63	16	1,733.3	4	4	2	2	0	Stairway	Flat	L shape	829,413,791
Case 64	6	646.7	2	3	1	2	0	Stairway	Flat	Linear shape	311,840,262
Case 65	9	985.3	2	5	1	2	1	Stairway	Flat	Linear shape	469,661,078
Case 66	5	545.9	1	5	0.5	2	0	Stairway	Flat	Linear shape	260,182,855
Case 67	7	772.0	2	4	1	2	1	Stairway	Flat	Linear shape	372,766,424
Case 68	7	892.3	2	4	1	2	1	Stairway	Flat	Linear shape	430,897,670
Case 69	24	2,749.9	4	6	2	2	0	Stairway	Flat	L shape	1,102,817,138
Case 70	41	4,336.9	4	11	2	2	3	Stairway	Flat	Linear shape	1,995,660,912
Case 71	26	3,055.6	3	10	1	3	4	Stairway	Flat	L shape	1,434,563,700
Case 72	20	2,174.1	2	11	0.5	4	2	Stairway	Flat	L shape	1,004,226,486

Case 73	22	2,281.6	3	10	1	3	8	Stairway	Flat	L shape	1,041,110,462
Case 74	31	3,355.5	3	11	1	3	2	Stairway	Flat	L shape	1,657,773,924
Case 75	34	3,646.7	3	12	1	3	2	Stairway	Flat	L shape	1,589,316,318
Case 76	22	2,396.5	2	12	0.5	4	2	Stairway	Flat	L shape	1,129,495,131
Case 77	24	2,556.7	4	6	2	2	0	Stairway	Flat	Linear shape	1,109,508,468
Case 78	15	1,623.1	4	4	2	2	1	Stairway	Flat	Linear shape	742,757,532
Case 79	16	1,733.3	4	4	2	2	0	Stairway	Flat	L shape	829,413,791
Case 80	6	646.7	2	3	1	2	0	Stairway	Flat	Linear shape	311,840,262
Case 81	9	985.3	2	5	1	2	1	Stairway	Flat	Linear shape	469,661,078
Case 82	5	545.9	1	5	0.5	2	0	Stairway	Flat	Linear shape	260,182,855
Case 83	7	772.0	2	4	1	2	1	Stairway	Flat	Linear shape	372,766,424
Case 84	7	892.3	2	4	1	2	1	Stairway	Flat	Linear shape	430,897,670
Case 85	21	2,390.9	6	4	2	3	3	Stairway	Flat	L shape	1,154,512,029
Case 86	24	2,749.9	4	6	2	2	0	Stairway	Flat	L shape	1,102,817,138
Case 87	41	4,336.9	4	11	2	2	3	Stairway	Flat	Linear shape	1,995,660,912
Case 88	26	3,055.6	3	10	1	3	4	Stairway	Flat	L shape	1,434,563,700
Case 89	20	2,174.1	2	11	0.5	4	2	Stairway	Flat	L shape	1,004,226,486
Case 90	20	2,220.4	2	11	0.5	4	2	Stairway	Flat	L shape	1,025,479,554
Case 91	22	2,281.6	3	10	1	3	8	Stairway	Flat	L shape	1,041,110,462

Case 92	31	3,355.5	3	11	1	3	2	Stairway	Flat	L shape	1,657,773,924
Case 93	34	3,646.7	3	12	1	3	2	Stairway	Flat	L shape	1,589,316,318
Case 94	22	2,396.5	2	12	0.5	4	2	Stairway	Flat	L shape	1,129,495,131
Case 95	37	3,980.7	3	13	1	3	2	Stairway	Flat	L shape	1,793,700,528
Case 96	34	3,660.9	3	12	1	3	2	Stairway	Flat	L shape	1,887,097,156

※ P1: Number of households, P2: Gross floor area, P3: Number of unit floor households, P4: Number of floors, P5: Number of elevators, P6: Number of household of unit floor per elevator, P7: Number of pilotis with household, Q1: Hallway type, Q2: Roof type, Q3: Building shape

초 록

건설 프로젝트의 완성에는 많은 시간과 자원이 소요되며, 이에 공사비 예측은 프로젝트의 전 단계에 걸쳐서 지속적으로 이루어져야 한다. 특히 초기단계 공사비 예측은 프로젝트의 성공 여부를 결정하고, 프로젝트가 진행될수록 공사비를 줄일 수 있는 가능성이 낮아지기 때문에 이러한 초기단계 공사비 예측은 매우 중요하다. 과거의 유사 사례를 바탕으로 새로운 문제를 해결하는 기법인 사례기반추론(case-based reasoning)은 인간의 문제해결 프로세스와 유사하며, 설득력 있고 정확한 해답을 빠르게 제시할 수 있으며, 유지관리가 쉽고, 사용할수록 사례 저장을 통해 정확도가 향상된다는 장점을 가지고 있으며, 이에 많은 연구자들이 사례기반추론을 공사비 예측에 적용하여 예측의 정확도를 향상시키기 위한 연구를 진행해 오고 있다.

사례기반추론 공사비 예측 모델의 예측 정확도는 조회(retrieving)와 보정(adaptation)영향을 받는다. 조회의 정확도는 사례를 표현하는 속성(attribute)의 수, 종류, 그리고 속성의 중요도를 나타내는 속성 가중치가 어떻게 적용되었느냐에 영향을 받게 된다. 그러나 기존의 사례기반 추론 방법들은 정성변수의 영향을 무시하거나 정성변수간의 차이를 반영하지 못하고 있으며, 이는 공사비 예측에 활용할 수 있는 속성의 수를 감소시키거나 유사사례의 유사도 순서를 왜곡하게 되며, 공사비 예측의 정확도를 감소시킬 수 있다. 또한 문제와 과거 사례가 완벽하게 일치할 수 없기 때문에, 과거 사례의 해결책(solution)은 새로운 상황에 맞도록 보정되어야 한다. 그러나 기존의 방법들은 추가적인 분석 혹은 데이터베이스, 학습 등이 필요하다는 단점이 있다. 또한 하나의 사례만 문제해결에 사용하는 것은 다른 유사사례들의 좋은 특성을 반영하지 못하게 되기 때문에, 여러 개의 사례를 통해 해결책을 도출해내는 방법(multiple case adaptation)이 활용되고 있으며, 사례 유사도를 가중치로 반영하는 가중 평균 방법

(weighted mean method)가 주로 사용된다. 그러나 이 방법은 유사사례 간의 가중치의 비가 상대적으로 적게 계산되어, 유사도 차이에 따른 보정 효과가 작다는 단점이 있다.

이러한 문제를 해결하기 위해, 본 연구는 글로벌 해를 찾을 수 있는 유전 알고리즘(genetic algorithm)을 활용하여 사례기반추론 코스트 모델에서 정성변수를 고려하기 위한 최적화 방법인 Dual-optimization을 제안하였다. 이는 속성 가중치와 함께, 정성변수의 속성값에 해당하는 random variable을 할당하고, 이를 함께 최적화함으로써 가중치 산정과 정성변수 정량화를 동시에 가능하게 하는 알고리즘이다. 또한 공사비 예측의 정확도를 더욱 향상시키기 위해, 사례 보정 문제를 해결하기 위한 조회 오차 보정 방법(retrieving error adaptation method)과 개선된 가중 평균 방법(improved weighted mean method)를 함께 제안하였다.

제안된 방법들을 군 병영생활관과 공공아파트 공사비 예측에 적용하고, 기존의 방법에 비해 정확도가 얼마나 향상되는지를 비교하여, 제안된 방법의 타당성을 검증하였다. 검증 결과 Dual-optimization 적용 시 공사비 예측의 정확도 및 안정성이 향상되며, 또한 이를 보정함으로써 예측 정확도를 더욱 향상시킬 수 있다는 것을 확인하였다. 이는 제안된 방법들이 공사비 예측에 충분히 활용 가능하다는 것을 의미한다.

본 연구는 기존의 방법들의 단점을 보완한 새로운 세 가지 방법을 제안하였다는 데에 그 의의가 있다. Dual-optimization 방법은 정성변수의 속성가중치 계산 및 정량화를 가능하게 하여 공사비 예측에 활용 가능한 속성의 수를 증가시킬 수 있으며, 조회 오차 보정 방법은 추가적인 분석이나 데이터베이스 없이 문제와 사례간의 차이에 대한 보정이 가능하다는 장점이 있다. 개선된 가중 평균 방법은 기존의 방법에 비해 더 유사한 사례의 영향을 증가시켜 더욱 정확한 해결책 도출을 가능하게 한다. 결과적으로 이를 적용한 새로운 GA-CBR 공사비 예측 모델은 더 정확한 공사비 예측을 가능하게 함으

로써, 건설 프로젝트의 참여자들에게 더 정확한 의사결정을 지원하는 도구로 활용될 수 있다. 또한 Dual-optimization은 범용적인 방법이기 때문에, 가중치 산정이 필요한 다른 분야의 문제에도 적용이 가능하다. 그러나 정성변수의 수가 많아질수록 계산시간이 오래 걸린다는 점은 사용성을 저해하는 요인이며, 이러한 문제는 컴퓨터 사이언스의 발전, 알고리즘의 개선 등을 통해 충분히 해결 가능할 것으로 기대된다.

주요어: 공사비 예측, 사례기반추론, 사례 조회, 사례 보정, 최적화, 유전 알고리즘

학 번: 2009-23032