S5-4

DATA MINING APPROACH TO PARAMETRIC COST ESTIMATE IN EARLY DESIGN STAGE AND ANALYTICAL CHARACTERIZATION ON OLAP (ON-LINE ANALYTICAL PROCESSING)

JaeHo Cho[1], HyunKyun Jung[2] and JaeYoul Chun[3]

[1] Ph. D. Candidate, Department of Architecture and Engineering, Dankook University, Korea
[2] Master Candidate, Department of Architecture and Engineering, Dankook University, Korea
[3] Professor, Department of Architecture and Civil Engineering, Dankook University, Korea
Correspond to cjhace@naver.com

ABSTRACT: A role of cost modeler is that of facilitating design process by the systematic application of cost factors so as to maintain sensible and economic relationships between cost, quantity, utility and appearance. These relationships help to achieve the client's requirements within an agreed budget. The purpose of this study is to develop a parametric cost estimating model for the early design stage by using the multi-dimensional system of OLAP (On-line Analytical Processing) based on the case of quantity data related to architectural design features. The parametric cost estimating models have been adopted to support decision making in the early design stage. These models typically use a similar instance or a pattern of historical case. In order to effectively use this type of data model, it is required to set data classification and prediction methods. One of the methods is to find the similar class in line with attribute selection measure in the multi-dimensional data model. Therefore, this research is to analyze the relevance attribute influenced by architectural design features with the subject of case-based quantity data used for the parametric cost estimating model. The relevance attributes can be analyzed by Analytical Characterization. It helps determine what attributes to be included in the OLAP multi-dimension.

Keywords: Data Mining, Classification, Parametric Cost Estimates, OLAP, Analytical Characterization

1. INTRODUCTION

Cost planning is a process in which the cost modeler determines the most suitable design alternative within an agreed budget by performing a cost evaluation based on various design attributes. To this end, the parametric cost estimating model was applied in the early design stage.

The parametric cost estimating method is able to support cost planning for various design attributes as it is possible to set impact factors related to design features as parameters(Kan Phaobunjong, 2002). The parametric estimating method uses pattern type of similar cases or of historical cases. In general, the cases used include geo-graphical design information, numerical information, BOQ(bill of quantity) and other types of information related to cost and quantity.

The parametric cost estimating method suggested in this study involves predicting the quantity that the cost modeler wants to determine based on the cases by utilizing the multi-dimensional information system (OLAP) to respond to queries according to impact factors (attributes) for each dimension.

The data structure of quantity cases is a relatively suitable model for a multi-dimensional analysis framework.

This research presents a concept of data classification approach in OLAP system to find similar case. The classification concept of data mining is to describe the data class or concept to find the collection of the distinguishing data, and the purpose is to use the model to predict the class object of quantity that is subject to building features conditions. The data classification relate to the attribute are useful for quantity prediction in the parametric cost estimating in the early design stage. And it can be easily be converted to logic rules.

2. METHOD OF RESEARCH AND SCOPE

The main driver used in the estimating model of this research is based on the OLAP multi-dimensional data cube. The cube makes the interactive querying in the mutual method for the desired attribute of user.

This research presents the data classification and prediction model for steel(rebar) ratio of RC Structure (Reinforced Concrete Structure) building that is related to the building attributes. For the items of quantity prediction in RC building in PA(Preliminary Appraisal) of the schematic design stage, there are quantity of forms, concrete, pouring, steel(rebar) and other minor works. Data analysis was made for the steel(rebar) per unit of concrete(m3) that constitute the building structure. Analysis of impact factors was limited to pure design factors related to change in quantity.

Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded. Measures of attribute relevance analysis can be used to help identify irrelevant or weakly relevant attributes that can be excluded from the concept description. It is referred to as Analytical Characterization or Analytical Comparison, respectively.

On the OLAP system, its attributes are applied to the sequence arranged in order of priority. In this case, it is possible to classification and forecast on the rule-based data model.

There have been many studies in machine learning, statistics, fuzzy and rough set theories, and so on, on attribute relevance analysis. The general idea behind attribute relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class or concept. Such measures include information gain, the Gini index, uncertainty, and correlation coefficients. This research introduces a method that integrates an information gain analysis technique (such as ID3 algorithm for learning decision tree) with dimension-based data analysis method. (A decision tree is flow-chart-like tree structure, where each node denotes a test on an attribute, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.)

3. PARAMETRIC COST ESTIMATING MODEL

In the early design stage, most cost estimating methods include the concept of parametric cost estimate. In the cost estimating research field, the parametric cost estimate is recognized as a very effective method in terms of its usability and efficiency.

The parametric cost estimating method is called the parameter estimating method or the cost estimating relationships method (CER's). The parameter estimating method is based on design variables and features that affect project size or scope. If the dependent variable of the CER's is set as quantity data, it is possible to predict construction cost with high reliability. In this case, since the prediction is conducted based on the quantity, basic drawings are required that are sufficient to calculate the quantity of each variable.

The recent trend in cost estimating technique is to use the object-based BIM (3D/4D). However, it is difficult to conduct object design in detail for cost estimating in the early design stage with the BIM design, and the design in detail is therefore skipped before the cost is calculated in connection with the statistical data module that is based on simple cases (for example, unit price rate: Cost/m2 GFA). The quantity prediction process is inevitably required for cost estimating with high accuracy in the early design stage.

Based on techniques that have been developed so far, it can be inferred that the OLAP technique in data mining includes classification and prediction functions of the CBR and ANN(Artificial Neural Network). Therefore, the OLAP system is a model that enables quantity prediction by making precise analysis of information on quantity case data. The OLAP technique can be used as an operation driver for quantity prediction of parametric cost estimating.

Such an operation driver enables the calculation of the final cost based on quantity prediction of the building element concerned after skipping unnecessary object design in detail in the early design stage.

4. STATUS OF EXISTING RESEARCH

Although the technical terms such as pattern recognition, class analysis and others are not usually used in the research of cost estimates ACE industry, we can easily recognize that a number of research have been conducted on the pattern recognition principle in data mining. For example, these researches have been continued in other name of pattern recognition that is Regression Analysis or Probability, CBR (Case-Based Reasoning), Neural Network or Fuzzy logic, etc.

The representative research status can be summarized as follow Table 1.

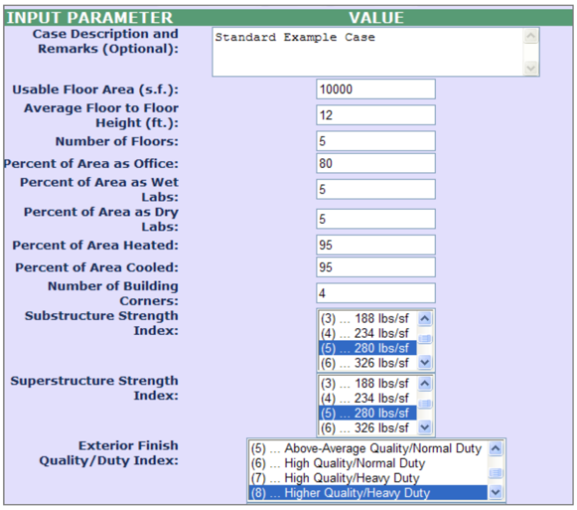
Table 1. Representative Research Status

Authors, (Year)	Estimating model	Attributes	Applied Data
I-Cheng Yeh, (1998)	Neural Network	Number of stories, Total height of building, Number of bays along frame, Typical bay length, seismic zone factor, live load, dead load, Compressive strength of concrete	RC structure Quantities Data
Wheaton and Simonton, (2005)	Regression (Statistics)	Number of stories, Absolute size, Number of units, Frame type, Year of construction	Office Cost Data
Elhag and Boussabaine, (1998)	Neural Network Model	Type of building primary, secondary school, Gross floor area, No. of stories, Project duration	School Cost Data
Sevgi Zeynep Doğan1, (2006)	CBR Model	Total area of the building, Ratio of the typical floor, total area of the building, Ratio of the footprint area, Number of floors , Type of overhang design, Foundation system, Type of floor structure , Location of the core	Residential Structure Cost Data
Ahmed A. Shaheen (2007)	Fuzzy Model	By work item (Optimistic, Most likely , Pessimistic)	Civil Cost Data

Looking into the existing research status, when users undertake cost estimating with the attributes, the applied concept is used in higher-level abstract. And in addition, the estimating process is determined by the data type (cost or quantity).

When looking under the point of view on parametric cost estimating, the major key would be to determine the type and dimension of such attribute. Therefore, from the several data mining approaches described in above, it is very important how to use the algorithm in certain model for data classification.

As shown in Figure 1 is basic model for the parametric cost estimates as presented by AACE(U.S.). The estimating process is automatically undertaken in accordance with the applicable attributes.



<Attributes>
1. Floor Area
2. Floor Height
3. Number of Floors
4. Percent of Area as Office
5. Number of Corners
6. Structural Strength/Duty Factors:
7. Structural Strength/Duty Factors
(Sub-Structure and Super-Structure)
In whole number from 0 to 10 - scalar indicates th e relative strength of the Sub-structure (caissons, slab, etc.) and Super-structure (above grade structure). Use the following rough guidelines based on the imposed floor loading in lbs per SF: 0 = 50 (light foot traffic, light roof, 1 = 96 (light office, heavy roof), 2 = 142 (heavy office, lab) , 3 = 188 (light mfg, mixed use) , 4 = 234 10 = 500 (v. heavy mfg, high live loads) , 5 = 280 (mixed mfg., warehouse)

Fig.1. Parametric Cost Estimating Model (AACE)

5. OLAP SYSTEM

Another way to express the definition of OLAP is "multi-dimensional information analysis". OLAP is defined as the "process in which the final user gets direct access to multi-dimensional information in order to analyze the information in an interactive way and utilize it for making a decision" (Choi, J.H., Park, S.J., 1996).

5.1 Structure of Case Data

The completed drawing documents include all of the information on quantities for geographical space, elements and works of building. Table 2 shows an example of information on the quantity case of concrete work in an office building.

The table 2 is shown in the form of total sum of elements for all standards of each work at the lowest level. Such information on quantity is previously summed up for floor and geographical space (underground floor, upper floor and foundation) before the aggregation of the information.

In addition, the quantities of elements of concrete, forms and rebar have been previously summed up for the standards of each element.

The rough hierarchical structure of quantity data can be understood in more detail by referring to the Galaxy Schema(Fact Constellation) in Figure 5.

Table 2. Information on quantity case of concrete work

Category	Concrete (m³)		Forms (m³)			Rebar (ton)		
Total Quantity	27,171.176		124,340.98			3,308.937		
Gross Floor Area	.69 m3/m2		3.156 m2/m2			.084 ton/m2		
Concrete			4.576 m2/m3			.122 ton/m3		
Element	Concrete (m³)	%	Forms (m³)	%	m2/m3	Rebar (ton)	%	ton/m3
Foundation	6,473.821	23.82	1,984.88	1.59	.307	674.896	20.39	.104
column	2,442.103	8.988	13,907.47	11.18	5.695	395.365	11.94	.162
Beam	2,682.943	9.874	10,152.33	8.16	3.784	812.046	24.54	.303
Slab	7,726.793	28.43	50,381.01	40.51	6.52	507.713	15.34	.066
Retaining Wall	7,057.254	25.97	41,889.91	33.69	5.936	818.979	24.75	.116
Staircase	327.941	1.207	2,152.62	1.73	6.564	40.684	1.23	.124
etc.	460.321	1.394	3,872.76	3.11	8.413	59.254	1.79	.129

5.2 Analysis of Quantity Case Data

The major task in this chapter is to explore which impact factor can be set as a multi-dimensional attribute in regard to quantity prediction.

Figure 2 and Figure 3 show the case of quantity for 20 office buildings in the 3D dispersion. Figure 2 indicates the total quantity cases of concrete and rebar in the dimensions of FA(Floor Area), number of story and unit rate.

The data information on the facility level (for element, geographical space and total sum) is expressed in the 3D space. Figure 3 describes the steel(rebar)/concrete ratio of the element for superstructure.

As shown in Figure 2 and Figure 3, it is difficult to clearly classify quantity information related to design information by using only single attributes in four dimensions. For example, data needs to be classified in more detail for the shape of building, facility service type, space attributes, structure type, and other special design features.

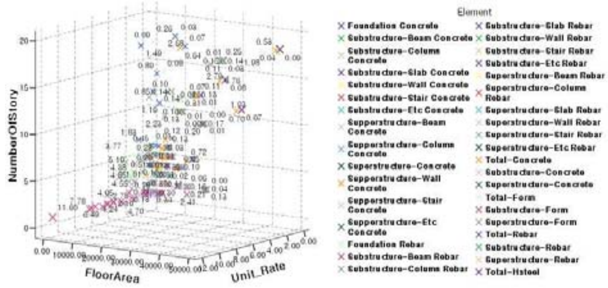


Fig.2. Quantity Case of Building Elements

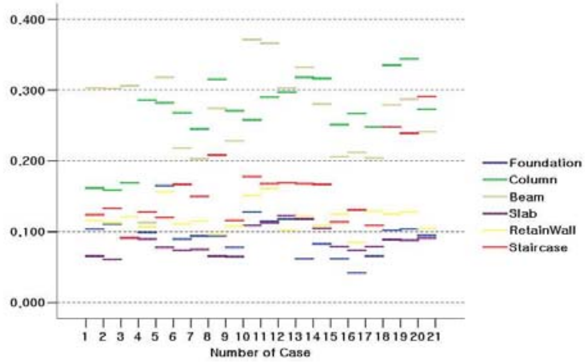


Fig.3. Steel(rebar)/Concrete Unit Ratio by Elements

5.3 Multi-dimensional Concept of the OLAP System

Data warehouse and the OLAP tool are based on a multi-dimensional data model. This model presents data as a type of data cube.

This chapter suggests how the data cube of the case quantity data can process modeling to become n-dimensional data.

The data cube is defined by dimension and fact. Dimension can be seen in the desired viewpoint for object, and each dimension has a dimension table that describes the details.

Figure 4 below shows the 4D data cuboid for quantity data. In the cuboid, the same terms are used as in the data cube. A cuboid lattice can be built from the set of given dimensions. Each cuboid shows data in the summary of other stages or "Group By" (summary of dimensions as other subsets).

Figure 4 indicates the cuboid in the dimensions of facility, shape, attribute, and composition as the dimension definition of the super-ordinate concept for impact factors (attributes) of quantity information.

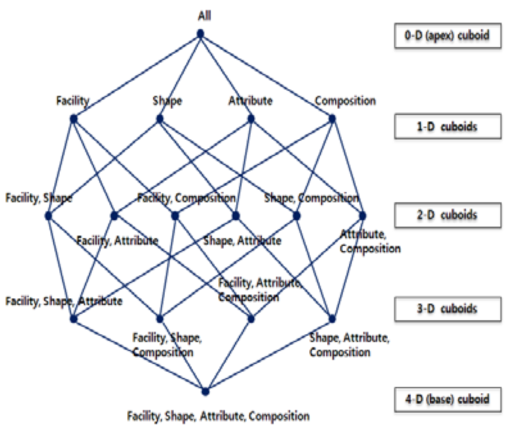


Fig.4. Cuboid Dimension Concept

5.4 Implementation of Data Schema

The most common data schema for data warehouse is basically a multi-dimensional model. Such a model can exist in the star schema, snowflake schema or galaxy schema types. For the utilization of each subject of the data information, several fact tables that share the dimension table may be required.

This kind of schema is considered as the set of star schema and is thus called "Galaxy Schema" or "Fact Constellation." Figure 5 shows the fact constellation schema suggested in this study. The schema indicates two fact tables such as space fact table and element fact table.

Each fact table shows unit rate information on the quantities of concrete, forms and rebar. Additionally, each fact table includes the ratios of constituent items in each quantity.

The fact constellation schema is used for modeling of a number of correlated subjects. This means that the information required for each cost estimating stage can be the information on quantity of geographical spaces (space fact table) or quantity of elements (elements fact table) depending on the level of progress in the design work.

Consequently, the suggested fact constellation schema is defined by distinguishing the requirements of the schematic design stage from those of the design development stage.

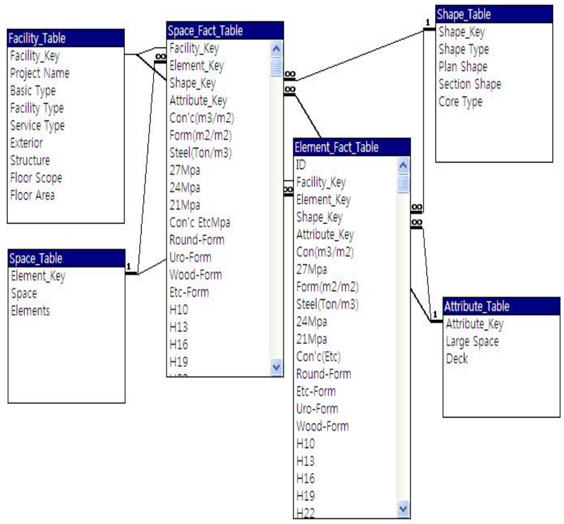


Fig.5. Fact Constellation Schema for Case Quantity Data

5.5 OLAP Reporting

The OLAP operation and query can be performed for various cases by using the implemented OLAP system. It is possible to inquire into the unit quantity of concrete against floor area for foundation, substructure and superstructure, depending on impact factors (attributes).

In addition, the ratios of form and rebar can be queried. The query of elements is also enabled. Such result reports are not provided in this study because of the limited space.

Figure 6 is a pie chart for impact factors of each dimension when the steel(rebar)/concrete ratio is queried for the superstructure of RC buildings.

Such multi-dimensional case analysis can provide statistical significance through the collection of data in a large quantity case. However, this study focuses on a query model for quantity case by the attributes.

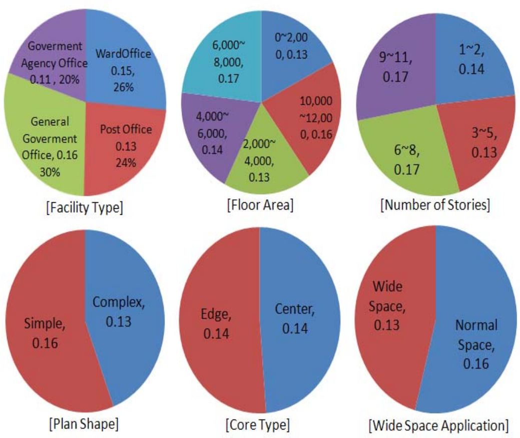


Fig.6. Analysis of Steel(rebar) /Concrete Unit Ratio

6. ANALYSIS CHARACTERIZATION

"What if certain attribute is included or excluded for class characterization and class comparison on certain data group, what should it be done?" If the attributes are used in relatively sufficient data, the clustering result may be good, but it may have complicated usability of the system with decline in capability.

Measures of attributes relevance analysis can be used to help identify irrelevant or weakly relevant attributes that can be excluded from the concept description process. In the event that too small attributes are used in the analysis process, the resulting minded descriptions may be incomplete.

As for the solution methods, it should be introduced to perform attribute (or dimension) relevance analysis in order filter out statistically irrelevant attributes, and retain or even rank the most relevant attributes for the descriptive mining task.

Therefore this research presents the information gain analysis technology that is used for decision making tree ID3 and C4.5 algorithm. This method removes the less informative attributes, collecting the more informative ones for use in concept description analysis.

The expected information gain calculation required for classifying the given sample is shown in the following formula (1).

Formula (1) : $I(s_1, s_2, \ldots, s_m) = I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$.

The Let S contain s_i samples of class C_i, for i=1, …, m. An arbitrary sample belongs to class C_i with probability s_i/s, where is the total number of samples in set S. An attribute A with values $\{a_1, a_2, \ldots a_v\}$ can be used to partition S into the subset $\{S_1, S_2, \ldots S_V\}$, where s_j contains those samples in S that have value a_j of A. Let s_j contain s_{ij} samples of class C_i. The expected information based on this partitioning by A is known as the entropy of A. It is the weighted average.

Formula (2) : $E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$.

The information gain obtained by this portioning on A is defined by $\text{Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A)$

7. CASE ANALYSIS

Table 3 displays the case collection consisted of data-tuple of the office buildings extracted from the DB. Data Class label attribute takes three values (high, middle, low) on the steel ratio for the RC building. In order to calculate the information gain on each attribute, it is measured expected information volume required in classifying the given sample by using Formula (1).

$$I(s_1, s_2, s_3) = I(12,11,4) = -\frac{12}{27}\log 3\frac{12}{27} - \frac{11}{27}\log 3\frac{11}{27} - \frac{4}{27}\log 3\frac{4}{27} = 0.919$$

Next is to calculate the entropy of each attribute category on the building service type.

Table 3. Training data tuples from the multidimensional database

Number	Building Type	Floor Area	Number of Floor	Plan Shape	Core Type	Large Space	Steel Ratio	Steel(Ton/con)
1	ward office building	4,000~6,000	Low	C	Eccentric Core	Yes	midium	15
2	ward office building	4,000~6,000	Low	B	Eccentric Core	No	high	16
3	ward office building	10,000~12,000	Low	C	Eccentric Core	Yes	midium	14
4	General Office	6,000~8,000	High	A	Center Core	Yes	high	17
5	General Office	0~2,000	Low	B	Center Core	Yes	high	16
6	**Post Office**	2,000~4,000	Low	B	Eccentric Core	Yes	high	16
7	**Post Office**	0~2,000	Low	B	Eccentric Core	Yes	low	12
8	ward office building	2,000~4,000	Low	C	*Eccentric Core*	Yes	midium	14
9	**Post Office**	6,000~8,000	Low	C	*Eccentric Core*	No	high	17
10	**Post Office**	2,000~4,000	Low	B	Center Core	Yes	low	10
11	General Office	10,000~12,000	High	B	Center Core	Yes	midium	15
12	ward office building	4,000~6,000	Low	C	Eccentric Core	No	midium	15
13	**Post Office**	0~2,000	Low	B	Eccentric Core	Yes	low	10
14	General Office	10,000~12,000	Low	B	Eccentric Core	Yes	high	18
15	**Post Office**	4,000~6,000	Low	C	Eccentric Core	No	midium	15
16	Special Agency	4,000~6,000	Low	C	Center Core	Yes	low	9
17	General Office	10,000~12,000	High	A	Center Core	Yes	high	17
18	Special Agency	4,000~6,000	Low	C	Eccentric Core	Yes	midium	13
19	General Office	10,000~12,000	Low	B	Center Core	Yes	high	17
20	General Office	10,000~12,000	High	A	Center Core	No	midium	15
21	General Office	4,000~6,000	Low	C	Eccentric Core	No	high	18
22	Special Agency	0~2,000	Low	B	Center Core	Yes	midium	13
23	General Office	2,000~4,000	Low	B	Center Core	Yes	midium	15
24	General Office	4,000~6,000	Low	C	Center Core	Yes	high	16
25	General Office	2,000~4,000	Low	B	Eccentric Core	Yes	high	17
26	Special Agency	4,000~6,000	Low	B	Center Core	Yes	midium	14
27	ward office building	2,000~4,000	Low	C	Eccentric Core	Yes	high	16

For Service Type = "Ward Office" $s_{11} = 2, s_{21} = 4, s_{31} = 0$,
$I(s_{11}, s_{21}, s_{31}) = 0.579$
For Service Type = "General Office" $s_{12} = 8, s_{22} = 3, s_{32} = 0$
$I(s_{12}, s_{22}, s_{32}) = 0.533$
For Service Type = "Post Office" $s_{13} = 2, s_{23} = 1, s_{33} = 3$
$I(s_{13}, s_{23}, s_{33}) = 0.921$
For Service Type = "Special Agency" $s_{14} = 2, s_{24} = 3, s_{34} = 3$
$I(s_{14}, s_{24}, s_{34}) = 0.512$

Using Equation (2), the expected information needed to classify a given sample if the samples are partitioned according to 'Service Type' is

$$E(\textit{Service Type}) = \frac{6}{27} I(s1_1, s2_1, s3_1) + \frac{14}{27} I(s1_2, s2_2, s3_2) + \frac{6}{27} I(s1_3, s2_3, s3_3) + \frac{4}{27} I(s1_4, s2_4, s1_4) = 0.63$$

Hence, the gain in information from such a portioning would be

$$\text{Gain (Service Type)} = I\,(s1, s2, s3) - E\,(\textit{Service Type}) = 0.29$$

With the same method, it can calculate for Gain (Floor Area)=0.15, Gain (Plan Shape)=0.05, Gain (Space Type)=0.04, Gain (Number of Floor)=0.03 and Gain (Core Type)=0.002. Since it has the gain on highest information for the building service type, the building service type would be selected as the most priority inspection attribute. Here, the threshold uses 0.03. Core Type is smaller than the information gain threshold that it is under the vulnerable relationship to remove this attribute.

As for the first in new node, it would have the label in the building service type, and in accordance with the size of each for the information gain of the attributes, the classification tree is generated. The final decision making tree generated by this algorithm is shown on the following Diagram. The attribute building service type has the highest information gain. Branches are grown for each type of building service. Therefore becomes a test attribute at the root node of the decision tree.

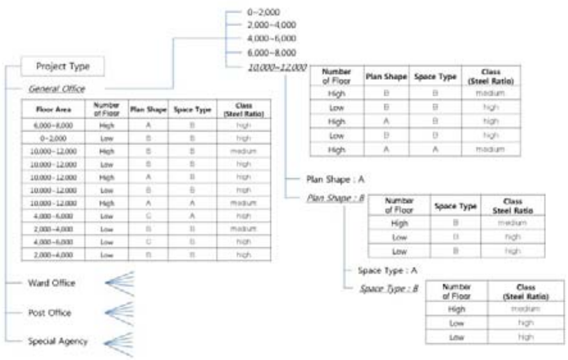


Fig.7. The decision tree for Steel(rebar) Ratio by information gain priority

8. CONCLUSIONS

In this study, the OLAP-based statistical quantity prediction method was suggested in regard to the parametric cost estimating method that is the subject of many studies for general purposes. The OLAP technique in data mining enables the user to make a decision based on a detailed comparison and analysis of target data at the multi-dimensional level.

The OLAP system used in this study can utilize impact factors(attributes) selectively that are required for cost estimating depending on progress in design.

Consequently, it enables cost estimating and cost planning with the help of quick query for various impact factors. In this test pilot, the case of steel(rebar)/concrete unit ratio prediction depending on impact factors was presented for a reinforced concrete building in the early design stage.

If the OLAP system is implemented by utilizing sufficient case data with the parametric cost estimating model, this is considered to be a highly reliable cost estimating method. However, if there is a problem related to the sparseness of the inquired case data, an advanced data mining technique should be applied such as the clustering method, classification, or association rule approach, etc.

In this case, data with similar cases are required to be clustered or summarized into an abstract concept (super-ordinate concept). This technique can be applied by using various cluster models or a classification approach in data mining.

This research has presented a case through the approach of analysis characteristic method for steel ratio change according to the design attribute.

The result of the data classification on the steel ratio is use to predict the steel quantity after calculating entire concrete quantity including slab, beam, girder, column, wall, stairs and others. The prediction model through the classification is available as the descriptive decision method in multi-level data model (OLAP).

ACKNOWLEDGMENTS

This research was supported by a grant (Assignment Number : 06 CIT A03) from the Construction Infra Technology Program funded by The Ministry of Land. Transport and Maritime Affairs and by a grant (Assignment Number : 2010-0025434) from National Research Foundation of Korea.

REFERENCES

[1] Christian Stoy, “Driver for Cost Estimating in Early Design”, ASCE, Journal of Construction Engineering and Management, Vol.134(1), pp. 26-32, 2008.
[2] Garold D. Oberlender, “Prediting Accuracy of Early Cost Estimaties Based on Estimate Quality”, ASCE, Journal of Construction Engineering and Management. Vol.127, No.3., 2001.
[3] Hao “Howard” Nie, “OLAP-Intergrated Project Cost Control and Manpower Analsis”, ASCE, Journal of Computing in Civil Engineering. Vol.21(3), pp. 26-32, 2007.
[4] J.P. Marques de Sa., “Pattern Recognition : concepts, methods, and applications”, Berlin ; New York: Springer, c. 2001.
[5] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers,2001.
[6] Kan Phaobunjong, “Parametric Cost Estimating Model for Conceptual Cost Estimating of Building, 2002. Construction Projects”, Doctor of Philosophy, The University of Texas.
[7] Richard J. Roiger, Michael W. Geatz, (2003) Data Mining A TUTORIAL-BASED PRIMER”, Addison-Wesley.
[8] Sevgi Zeynep Dogan, (2008) Using Decision Trees for Determining Attribute Weights in a Case-Based Model of Early Cost Prediction , ASCE, Journal of Construction Engineering and Management, Vol.134, No.2.