MINING OF CONCEPTUAL COST ESTIMATION KNOWLEDGE WITH A NEURO FUZZY SYSTEM

Wen-der Yu and Yu-ru Lee

Institute of Construction Management, Chung Hua University, Taiwan

Abstract: Conceptual cost estimation during the early stage of a construction project plays important role for feasibility analysis and project planning. Traditional approaches rely heavily on experienced engineers, and may cause loss of conceptual estimation knowledge of the firm. This paper proposes a method integrates a previous developed conceptual cost estimation method (PIREM) with the ANFIS neuro-fuzzy system for mining of cost estimation data. A case study of residential building projects in Mainland China is conducted to test the proposed method. The testing results show that the proposed method does not only achieve high system accuracy, but also provide many features desirable for estimators such as explicit fuzzy decision rules and graphical sensitivity analysis presentation.

Keywords: Conceptual cost estimation; Neuro-fuzzy; Data mining, KDD; China

1. INTRODUCTION

The conceptual cost estimation during the engineering planning is important for successful execution of a construction project, since the main structural systems, major construction methods, and most construction materials are determined in that stage. However, due to the lack of detail design information during the planning phase, accurate cost estimation is hard to obtain even for the professional estimators. It was found that the estimators with more estimating experience can do better in this job than who with less. The emerging development of modern artificial intelligence (AI) techniques, such as neuro-fuzzy systems, the aforementioned estimating experience/knowledge can be acquired by learning from historical examples, so that accurate estimation (compared with the detail estimation) could be obtained with very limited available project information. Unfortunately, an essential difficulty is facing the traditional conceptual cost estimation if the knowledge-based approaches are to be adopted—the unit prices of the cost items are variable in the marketplace, so that the estimation knowledge learned previously may not be readily applicable in the future projects.

In this paper, the PIREM (Principal items ratio estimation method) [1] approach is integrated with a neuro-fuzzy system, ANFIS [2], to perform data mining (DM) function in the knowledge discovery process of conceptual cost estimation. An application of the proposed DM method for cost estimation of building construction projects in People's Republic of China (PRC) is selected for demonstration of the proposed method in order to meet the needs caused by more and more construction projects invested by the Taiwan's businessmen in PRC in the past decade. As the cost estimation system of PRC (so-called "fixed price system [3]") is different from the cost estimation system in Taiwan (i.e., "bill of quantities", BOQ system), investors from Taiwan and other countries are unable to obtain accurate conceptual cost estimates for their projects, especially for the first-time investors. In order to conquer the difficulty, knowledge discovery in databases (KDD) [4] techniques are employed to mine the cost estimation knowledge from historical cost data of previous construction projects. The historical cost data are collected from sample projects in the publications published by the Ministry of Construction of PRC [5,6]. Totally, 114 building construction projects were collected and analyzed. The quantities and their unit price information of every cost item are surveyed and calculated separately in PIREM. The ANFIS neuro fuzzy system is adopted to capture the relationships between the influential attributes and the construction cost. The relationships acquired by ANFIS are stored in forms of fuzzy IF-THEN rules, so that the domain experts can visualize and verify the cost estimation knowledge explicitly. With the aid of the proposed approach, the barrier caused by different cost estimation system can be overcome. The testing results show that the cost estimation accuracy can be up to 90.01%, which is considered acceptable during the project conceptual planning phase.

The paper is organized in the following manner. In the second section, the previously developed PIREM approach for conceptual cost estimation is reviewed first to provide background knowledge. Then, process for mining of historic cost estimation data is briefly discussed. Fourthly, the ANFIS neuro-fuzzy system as a technique for data mining is introduced. In the fifth section, application of the proposed approach to cost estimation of building projects in Mainland China is presented. Both data mining and system testing are presented and discussed. Finally, the findings of the research are concluded. Difficulties encountered in KDD of conceptual cost estimation, especially for cost estimation of projects in PRC, are discussed. Directions of future research are also suggested at the end of this paper.

2. PRINCIPAL ITEMS RATIO ESTIMATION METHOD (PIREM)

Among the many conceptual estimating methods, parametric cost estimating has been widely adopted in industry for the economic feasibility analysis in the early stage of a construction project. The parametric cost estimating takes important influential parameters as inputs, such as the floor area, cubic volume, bay width, etc. By statistic regression or other mapping schemes, the relations between the estimates and the influential parameters are established. The cost estimates of new projects are obtained by mapping inputs of parameter values based on pre-determined mathematic relation [1].

2.1 Problems facing traditional methods

The essential problem for all parametric cost estimating approaches is that the unit prices of cost items fluctuate as time passes. Thus, construction cost estimates obtained based on previous estimation experiences may be incorrect due to unit price variation in the marketplace. The key problem to this result is the use of "activity cost" as the measure of estimates during analogizing process. However, the "activity cost" mixes up the two elements—quantity and unit price—of a construction cost item. Even though some approaches introduced the "overall price index" as a parameter for adjusting construction costs [7], the price fluctuations for individual cost items are not varying proportionally and simultaneously. Adjusting the unit prices of all cost items with a "overall price index" is not realistic. The PIREM proposes a better way is to divide the two elements of the cost item and handle them separately. The quantities of a cost item are mainly affected by the factors of construction method adopted, such as the dimensions of structural design, the method of construction, the conditions of site environment, etc. Thus, they won't change as long as the same facility is constructed by the same method under the same environmental conditions. On the other hand, the unit price of the cost item may vary from time to time. It is reasonable to employ the most updated unit prices for the cost items in order to reflect the prices of marketplace.

Another problem is that a construction project usually consists of hundreds or even thousands of cost items. It is very expensive and time consuming to obtain the quantity estimates and unit prices of all items in a construction project. In order to resolve this problem, the PIREM adopts the Pareto Optimum Criterion (or namely, "80/20 Principle") to simplify the estimation process.

2.2 Model of PIREM

With the Pareto Optimum Criterion, only the top 20% cost items is selected as "*Principal Items* (PI)". The summation of the principal cost items gives the *Cost of Principal Items* (*PIC*). The ratio of *PIC* over the overall cost is defined as the *Principal Item Ratio* (*PIR* or *p*). The value of *p* can be calculated by the following equation.

$$p^s = \frac{\sum_{j=1}^{l} UP_j^s \cdot Q_j^s}{\sum_{i=1}^{n} UP_i^s \cdot Q_i^s} = \frac{PIC^s}{OC^s}, \qquad (1)$$

where the numerator is the summation of the costs of all *principal items*, while the denominator is the summation of the costs of all cost items in a project. The super script *s* upon all parameters stands for the s^{th} historical example. UP_i means unit price and Q_i means quantity of the i^{th} cost item, and so forth. *OC* stands for overall cost. The ratio *p* obtained in Eq. (1) is called *Principal Item Ratio* (*PIR*) as defined above.

It is found by analyzing *PIR*'s obtained from historical cost estimation data that the *PIR* of a specific type of construction project usually keeps constant with very small variation [8]. Therefore, given the *PIC*, the *OC* of the new project can be recovered by the following equation.

$$OC^r = \frac{PIC^r}{\bar{p}}, \qquad (2)$$

where OC^r is the estimate of overall cost for the new (r^{th}) project. $\bar{p}$ is the average *PIR* calculated from previous projects. PIC^r is the cost of principal items of the new (r^{th}) project.

In Eq. (1) and (2), the current unit prices (*UP*'s) for the new project should be surveyed from the marketplace at the moment of estimation. While, the quantity, *Q*, and average *PIR*, $\bar{p}$, have to be determined by parametric estimating methods based on the pre-determined mathematic relations. The mathematic relations between input parameters and the output estimate values are usually nonlinear. There are some approaches available for establishing such relations including nonlinear regression, artificial neural networks, and case-based reasoning. Among those, the nonlinear regression method may encounter severe convergence difficulty while dealing with more than two variables. The artificial neural network approach is good at nonlinear mapping. However, the mathematic relationships established are stored in a "black box" thus limits the value of knowledge learned. The estimation error of case-based reasoning method is still high [7], which may not be accurate enough for decision-making during conceptual planning stage. In the proposed system, a neuro-fuzzy soft computing technique, ANFIS [2], is adopted to achieve more accurate estimations and provide human understandable knowledge for the estimators.

3. DATA MINING PROCESS

Data mining was defined as the application of automated knowledge acquisition methods for generation of useful knowledge via organization and analysis of raw data [9]. The procedure for data mining implementation consists of five steps [10]: (1) objective determination; (2) data preparation; (3) data transformation; (4) data mining; and (5) result analysis. There are some information techniques available for data mining implementation such as symbolic learning, case-based reasoning, and artificial neural networks. Two key issues for data mining are: (1) the accuracy of knowledge acquisition, and (2) the format knowledge representation. In this paper, a neuro-fuzzy system named ANFIS is adopted for mining of cost estimation data primarily due to the excellent learning ability and explicit knowledge representation of neuro-fuzzy systems.

Descriptions of the process for mining of historical cost estimation data are follows.

(1) Objective determination—the goal of data mining for cost estimation is to acquire the underlying knowledge embedded in the historical cost estimation data. Two primary objectives are providing accurate estimate of construction cost with a set of given parameters and the insight (inference process) of cost estimation.

(2) Data preparation—the “raw data” for cost estimation are collected from former projects and cleaned via a data pre-treatment process.

(3) Data transformation—the data transformation consists of two tasks: quantification of qualitative parameters and data normalization. This step is performed to prepare data for ready to use by the data mining techniques.

(4) Data mining—in this paper, the data mining process is performed by ANFIS neuro-fuzzy system. A commercial software — Matlab™ Fuzzy Logic Toolbox® — is adopted for mining of the historical cost estimation data. The raw data are divided into two sets: the training set and testing set. The grid partition scheme is used for fuzzy partition of input parameters. A hybrid-learning rule combining steepest descent scheme and least-square-error scheme is adopted for training of the neuro-fuzzy system.

(5) Result analysis—after training, the fuzzy decision rules are extracted from the historical cost estimation data for each training set. The rules can be used for system testing. They can be evaluated by the domain experts.

4. ANFIS NEURO-FUZZY SYSTEM

Neuro-fuzzy system is a branch of artificial intelligence (AI), which combines the merits of artificial neural networks (ANN) and fuzzy inference systems (FIS). A basic structure of an FIS is comprised of three components: (1) a rule base, which stores a bunch of fuzzy if-then rules; (2) a database, which defines the membership functions used in the fuzzy rules; and (3) a reasoning mechanism, which performs fuzzy inference upon the rules and given facts to derive a reasoning output or conclusion. There are three major types of FIS’s widely adopted for management and engineering applications:

(1) Mamdani FIS—a general type of FIS that adopts “max” and “algebraic product” for fuzzy T-norm and T-conorm operations. A typical fuzzy if-then rule for Mamdani FIS is shown in Eq. (1).

$$R^k : \text{If } x_1 \text{ is } A_1^k \text{ and} \ldots \text{and } x_p \text{ is } A_p^k, \text{Then } y \text{ is } B^k, \quad (1)$$

where R^k means the k^{th} fuzzy rule; A_i^k and B^k represent fuzzy linguistic variables; $\bar{x} = (x_1, x_2, ..., x_p)^T \subset \Re^p$ and $y \subset \Re$ are the inputs and output of the k^{th} fuzzy rule.

(2) Sugeno FIS—a special type of FIS that adopts a crisp function in the consequence of a fuzzy decision rule. A typical fuzzy decision rule for Sugeno FIS is shown in Eq. (2).

$$R^k : \text{If } x_1 \text{ is } A_1^k \text{ and} \ldots \text{and } x_p \text{ is } A_p^k, \text{Then } y \text{ is } f_{\text{k}}(x_1, x_2, ..., x_p), \quad (2)$$

where R^k, A_i^k, B^k, $\bar{x} = (x_1, x_2, ..., x_p)^T$, and y are defined similarly as in Eq. (1), while $f_{\text{k}}(x_1, x_2, ..., x_p)$ is a polynomial taking on x_1, x_2, …, x_{p} and is used to define the consequence of a fuzzy decision rule.

(3) Tsukamoto FIS—also a special type of FIS that adopts a monotonical membership functions in the consequent part of a fuzzy decision rule. A typical fuzzy decision rule for Tsukamoto FIS is shown in Eq. (3).

$$R^k : \text{If } x_1 \text{ is } A_1^k \text{ and} \ldots \text{and } x_p \text{ is } A_p^k, \text{Then } y \text{ is } C_{\text{k}}. \quad (3)$$

where R^k, A_i^k, B^k, $\bar{x} = (x_1, x_2, ..., x_p)^T$, and y are defined similarly as in Eq. (1), while C_{k} are monotonical membership functions to describe the consequent part of a fuzzy decision rule.

While applying to data mining, the three FIS’s mentioned above need to be equipped with “learning abilities” so that they are able to “mine” knowledge from raw data. Three popular learning schemes for constructing a neuro-fuzzy system are: (1) FALCON proposed by Lin and Lee [11], (2) ANFIS proposed by Jang [2], and (3) Back-propagation fuzzy system proposed by

Wang and Mendel [12]. In this paper, the ANFIS is adopted for construction of a Sugeno FIS.

5. APPLICATION TO RESIDENTIAL CONSTRUCTION PROJECTS IN CHINA

In this section, the proposed method is applied to cost estimation of residential building construction projects in Mainland China. The application is challenging in construction cost estimation due to the wide range of construction locations and variety of construction firms.

5.1 Fixed-price system in China

The "fixed-price system" is an outgrowth of the planned economy system of socialism, which aims at unifying the quantities and prices of a specific construction cost item by the government, so that the national economy can be under control [3]. In practice, the estimators of contractors usually prepare a "main resources list (MRL)" of the project while they submit their bids. Such MRL is very similar to the *PI* in PIREM described previously in this paper. With MRL, the managers are able to visualize cost profile of the construction project.

5.2 Why PIREM for construction cost estimation in China?

PIREM has been successfully applied in the public construction cost estimations in Taiwan [8]. The aptly separation of quantity and unit price of the cost item has demonstrated its capability in reflecting the most updated currency in marketplace. The "fixed price system" of Mainland China also separates the quantity with unit price determined by the government—the "fixed price", thus shows great similarity with the scheme of PIREM. The other reason for adopting PIREM is the wide range and great variety of China's construction market. Since China is one of the largest countries in the world, it is difficult for a contractor to develop specific estimation system for every province or city. PIREM provides a universal conceptual cost estimation method for contractors who perform projects in different provinces or cities in China.

5.3 Historical databases

In order to acquire the cost estimation knowledge, historical data of 110 high-rise residential construction projects from China are collected from a government publication [5,6]. The location distribution of the 110 projects is shown in Figure 1. The 110 projects are selected from 25 provinces/cities, where the top five provinces/cities are Beijing City, Fujian Province, Hebei Province, Hunan Province, Sichuan Province, and Shaanxi Province.

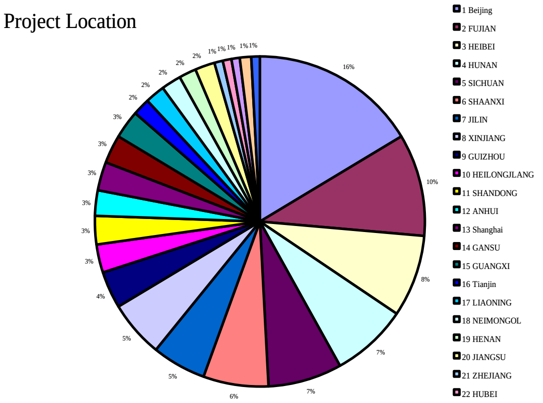

Figure 1. Distribution of project locations

The cost volume distribution of the110 projects is shown in Figure 2. In Figure 2, it shows that the construction costs of the historical residential projects are uniformly distributed.

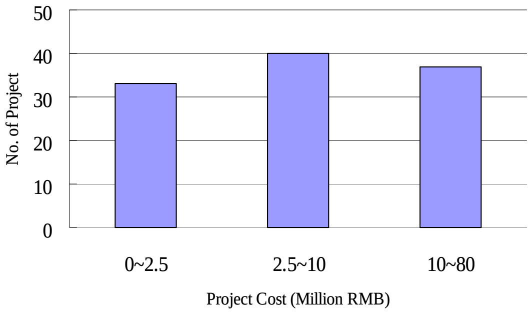

Figure 2. Distribution of construction costs

By statistic analysis, the MRL of the 110 residential construction projects consists of 10 elements: labor, cement, rebar, gravel, sand, brick, structural steel, tile, wood, and glazing, which will be considered for principal items (PI) of PIREM.

5.4 Verification of ANFIS data mining

To verify the nonlinear mapping of ANFIS, historical data were collected and divided into two sets: (1) 110 data sets collected from reference are used for training, and (2) the other 10 data sets collected from reference are used for testing. The training error is controlled within the error goal of 10%. The maximum epochs for training are controlled within 1000. The training and testing are performed for all principal cost items. The testing errors are within 10% (9.99%). The system accuracy is generally satisfied, since the acceptable error for a conceptual estimation system usually ranges from 15%~20% [13]. The results for the 10 testing sets are shown in Table 1, where the foundation types consist of: (1) simple foundation (value=1); (2) continuous foundation (value=2); (3) shaft foundation (value=3); (4) box foundation (value=4); (5) PC pile (value=5). Moreover, the structural types include: (1) brick (value=1); (2) RC frame (value=3); (3) shear wall (value=5).

Table 1. Results for testing sets

No.	PIC*	P%	Actual cost*	Estim. cost*	Acc. %
1	290	47.70	556	607	90.88
2	564	51.70	1057	1091	96.75
3	292	45.30	626	645	97.08
4	491	42.10	1015	1166	85.19
5	324	46.30	610	700	85.37
6	503	42.30	1018	1189	83.13
7	232	53.60	462	433	93.65
8	459	45.90	1114	1000	89.75
9	488	44.70	1224	1091	89.20
10	596	45.60	1179	1306	89.15
Average accuracy %					90.01

* RMB/M^2

5.5 Influence of price fluctuations

In order to assess the influence of price fluctuations in the market, two scenarios are designed for sensitivity analysis: (1) all unit prices of principal items are increasing while the non-principal items are decreasing; (2) all unit prices of principal items are decreasing while the non-principal items are increasing. Each of the above scenarios is analyzed under three ranges of price variation, that is: (1) unit prices randomly varying from 0~5%; (2) unit prices randomly varying from 5~10%; and (3) unit prices randomly varying from 10~15%. The price variation ranges is compared with the price fluctuations observed from the market shown in Table 2. It's found from Table 2 that the maximum price variation of China's market is below 5%. Therefore, the sensitivity analysis should cover all possible price fluctuation ranges that may happen in the near future.

Table 2. Price fluctuations of China, 1996~2001 [14]

Year	Consu. index	Retail index	Manu. index	Matl . index	Fixed prop. index
1996	100.00	100.00	100.00	100.00	100.00
1997	102.81	100.79	99.68	101.30	101.69
1998	101.98	98.17	95.60	97.02	101.47
1999	100.55	95.24	93.29	93.83	101.07
2000	100.95	93.81	95.92	98.62	102.20
2001	101.65	93.07	94.68	98.40	102.59
Fluct.	1.56%	4.34%	4.40%	2.79%	1.77%

Table 3. Testing of unit price fluctuation

Scenario	Unit price variation	Ave. accuracy.	Error increase
Scenario I ([1]PUP↑, [2]NUP↓)	0%	90.01%	0.00%
	0~5%	90.98%	↑ 0.97%
	5~10%	88.08%	↑ 1.93%
Scenario II ([1]PUP↓, [2]NUP↑)	0%	90.01%	0.00%
	0~5%	91.45%	↓ 1.44%
	5~10%	90.09%	↑ 0.08%

[1]PUP: Unit prices of principal items

[2]NUP: Unit prices of non-principal items

The proposed system is also tested to view the influence of unit price fluctuation on its estimation accuracy. The testing results are shown in Table 3. Where, under the worst scenarios, the highest error is still below 20%, the margin of maximum allowable error for a conceptual estimation system [13].

5.6 Sensitivity analysis

One of the most valuable features for data mining is the graphical presentation of the minded patterns or knowledge. For cost estimation, user always wants to know the most sensitive factors affecting the construction cost. In this regards, sensitivity analysis of various influential attributes on overall construction cost is very useful for value engineering and best alternative selection. Figure 3 and 4 show examples of sensitivity analyses, where the red color indicates areas that are potentially costly and should be avoided. Moreover, sensitivity analysis can also help users identify cost sensitive attributes. For example, the foundation type is more cost-sensitive than structure type and floor area in Figure 3; the number of stories is more cost-sensitive than structure type in Figure 4.

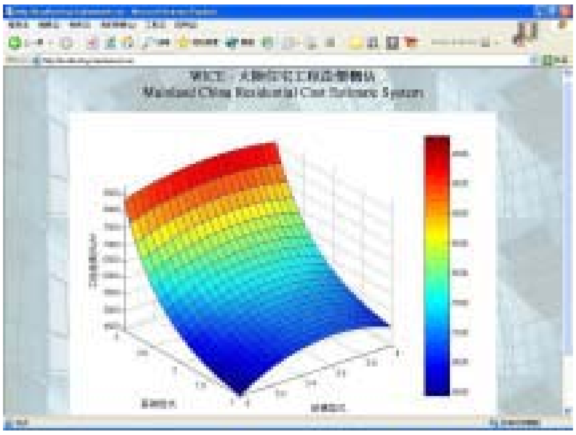

Figure 3. Sensitivity Analysis—Foundation type vs. Structure type

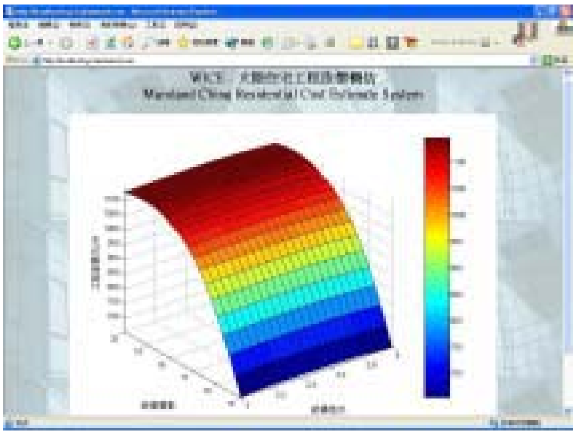

Figure 4. Sensitivity Analysis—No. of stories vs. Structure type

5.7 Fuzzy rule base

The knowledge mined by ANFIS is stored in the fuzzy rule base such as the one shown in Figure 5. The fuzzy rule base contains a set of fuzzy decision rules. Every fuzzy decision rule consists of a set of fuzzy linguistic terms for expressing values of every attribute in the precondition part; it also contains a set of fuzzy linguistic terms for the single output in the consequence part. Every fuzzy linguistic term is coupled with a fuzzy membership function. The fuzzy decision rules can be visualized and evaluated manually by the domain experts.

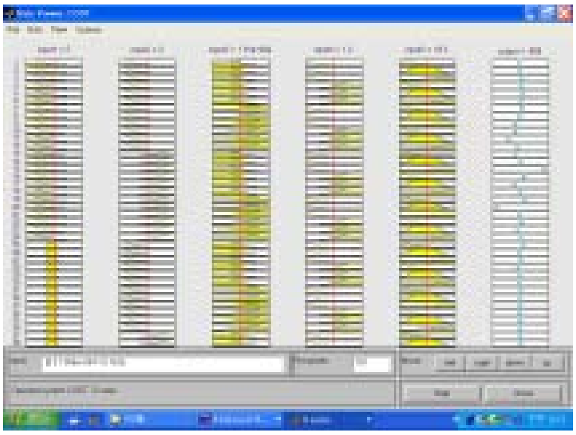

Figure5. Fuzzy rule base

6 CONCLUSIONS AND RECOMMENDATIONS

This paper presents a data mining method for mining of construction cost estimation knowledge. The proposed method integrates a conceptual cost estimation method (PIREM) with a neuro-fuzzy soft computing data mining technology (ANFIS) to provide desirable features for construction cost estimation. While applying to the geometrically huge and geographically complex construction market such as China, the proposed method demonstrates its outstanding capabilities in knowledge discovery. It provides not only the accurate estimation results but also the visual knowledge representations that are useful and desirable for the users. A case study of residential building construction projects in Mainland China is demonstrated to test the proposed method. The testing results show that the proposed method can still achieve high estimation accuracy up to 90.01% even for the great variety market such as China.

Some future directions for research can be pursued: (1) application of proposed method to other types of projects, such as industrial projects logistic inventory projects, infrastructure projects, etc.; (2) integration of proposed method with construction planning and scheduling systems to expedite project execution; (3) development of a more intelligent data mining technique to improve the estimation accuracy.

7. REFERENCES

[1] Yu, W. D., and Lai, C. C., “WICE: A Web-based Intelligent Cost Estimator for Real-time Decision Support,” *Proceedings of IEEE/WIC International Conference on Web Intelligence (WI'03)*, Oct. 13~16, Session 10B, October 13-16, 2003, Lord Nelson Hotel, Halifax, Canada, pp. 646~649, 2003.

[2] Jang, J. S., “ANFIS: Adaptive-network-based fuzzy inference system,” *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 23, No. 3, 1993, pp. 665-685.

[3] Guo, J. J., *The Fixed Price Costing and Conceptual Budgeting*, Northern Jiaotong University, Beijing, P.R.C., 2003. (in Chinese)

[4] Fayyad, U., and Uthurusamy, R., “Data mining and knowledge discovery in databases,” *Commun. ACM,* Vol. 39, pp. 24–27, 1996.

[5] Standardization and Fixed Pricing Institute, *The Economic Indexes of Private Construction*, Published by Ministry of Construction, Beijing, P.R.C., 1996. (in Chinese)

[6] Standardization and Fixed Pricing Institute, *The Economic Indexes of Private Construction*, Published by Ministry of Construction, Beijing, P.R.C., 2002. (in Chinese)

[7] Yu, J. S., “Developing building cost estimating system using case-based reasoning approach,” *Master Thesis*, Department of Civil Engineering, National Central University, Chungli, Taiwan, R.O.C., 2001. (in Chinese)

[8] W. D. Yu, and J. B. Yang, *Final Report on the Development of a Neuro-Fuzzy Knowledge-based System for Construction Conceptual Estimation*, CECI, Taipei, Taiwan, 2002. (in Chinese)

[9] Michalski, R. S., Bratko, I., and Kubat, M., *Machine Learning and Data Mining: Methods and Applications*, Wiley, NY, 1998.

[10] Leu, S. S., Chen, C. N., and Chang S. L., “Data mining for tunnel support stability: neural network approach,” Journal of Automation in Construction, 10 (2001), pp. 429-441, 2001.

[11] Lin, C. T., and Lee, C. S. G., “Neural-network-based fuzzy logic control and decision system,” *IEEE Transactions on Computers*, Vol. 40, No. 12, pp. 1320-1336, 1991.

[12] Wang, L. X., and Mendel, J. M., “Back-propagation fuzzy systems as nonlinear dynamic system identifiers,” *IEEE Trans. on Neural Networks*, Vol. 3, pp. 807-814, 1992.

[13] Jong, S., “Object-oriented cost estimation system,” Master *Thesis*, Department of Civil Engineering, National Taiwan University, Taiwan, R.O.C., 1992. (in Chinese)

[14] *Statistical Indicators*, National Bureau of Statistics of China, http://www.stats.gov.cn/, 2004.