2012 International Conference on Industrial and Intelligent Information (ICIII 2012)
IPCSIT vol.31 (2012) © (2012) IACSIT Press, Singapore

Application of Case-Based Reasoning in cost estimation of drilling wells

Hossein Shams Mianaei[1], Seyed Hossein Iranmanesh[2], Hamid Akbari[3]

Industrial Engineering Department, College of Engineering, University of Tehran, Tehran, Iran
Petropars Company, South Pars Phase 12, Assistant Professor of IIES, Tehran, Iran

Abstract. Costs estimation of drilling for new wells is one of the main concerns of active companies in this field. Since some of features of drilling can not be expressed quantitatively and there are many qualitative features in data of the available wells. So a method should be applied to use these data to estimate the desired and ideal output of project manager. The case based reasoning (CBR) method covers the qualitative data with regard to its nature. Using CBR method which is created based on the viewpoint of using presented solutions for previous solved problems in order to solve new similar problem save time and therefore speed of drilling is increased which it is very important, as regards available estimation methods spend the much time to estimate output which is cost of drilling wells in this case study. The results show that obtained estimation accuracy is high.

Keywords: Case Based Reasoning (CBR); Cost Estimation

1. Introduction

Costs estimation accuracy is an important factor in project success. Scientific methods should be employed during project planning and design to increase costs estimation accuracy. The cost estimation model of case-based reasoning (CBR), which in preliminary stages estimates the costs with minimum project information, is useful in preliminary design stage of the drilling wells project.

CBR method uses earned experiences in solving past problems as a guide for solving new problems. Solving problem by using CBR method is done in one cycle. When a new problem is presented, its conditions is compared with status of previous solved problems and using mechanisms of similarity, the most similar previous cases are retrieved. Then, the retrieved cases are applied to provide a solution for the new problem and subsequently proposed solution is provided. If required, the proposed solution will be revised according to the position of the new problem and finally, new case (i.e. the considered problem and its solution) is retained in case base for future usage.

2. Literature review

In this section, we briefly review the prior fundamental studies on CBR in cost estimation. In late 1980s, a new approach, called expert systems, was introduced to estimate cost. However, the use of expert systems did not reach its maximum potential. Therefore, CBR systems were proposed as an alternative for expert system in cost estimation. For example, Perera and Watson [1] proposed a prototype system, NIRMANI, based on CBR in order to support design and cost estimation. During design stage, 80% of the cost of a product is planned. In this field, Duverlie and Castelain [2] showed the application of parametric methods and CBR (induction) method for cost estimation in design phase of mechanical production and more particularly, the production of pistons for railway diesel engines. Kim et al. [3] examined the performance of three cost estimation models. The examinations are based on multivariate regression analysis (MRA), neural

[1] *E-mail address*: h.shams.m@gmail
[2] *E-mail address*: hiranmanesh@ut.ac.ir
[3] *E-mail address*: akbary@ppars.com

networks (NNs), and CBR from the data of 530 historical costs. As far as we know, up to 2010, there was no model which could be applied directly in forecasting manufacturing costs. Research of Chang et al. [4] made the first attempt for development of a hybrid system by integrated CBR and Artificial Neural Networks (ANN) as a forecasting model of Product Unit Cost (PUC) in Mobile Phone Company. The proposed model in their research was compared with other five models. Their findings indicated that MAPE value of the proposed model was smallest.

3. The proposed CBR model for cost estimation of drilling wells

This study develops a CBR model for cost estimation of drilling wells based on features of the project of drilling well (case). Fig. 1 shows the process of proposed CBR model.

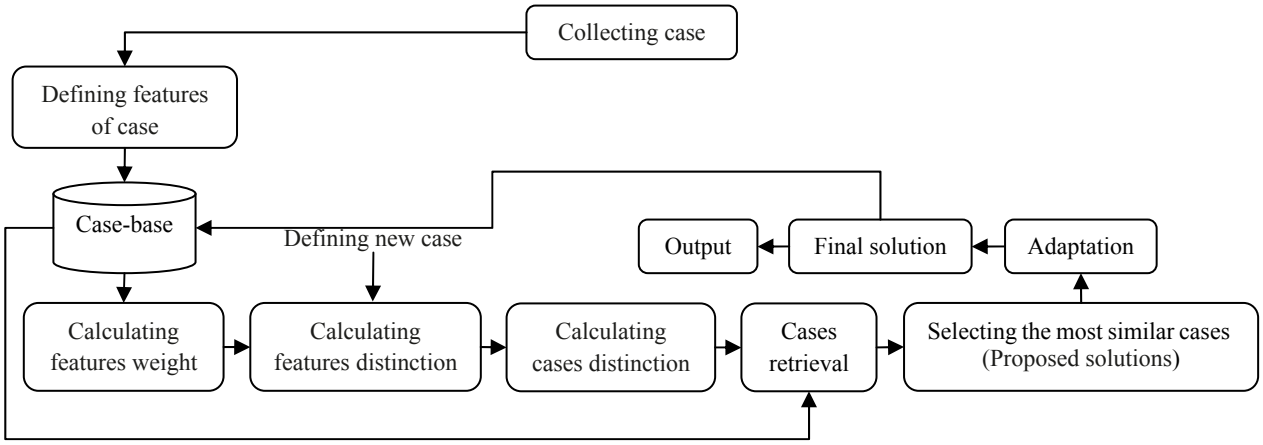

Fig. 1: the process of proposed CBR model

In the first stage, features and performance indices of project and their scale are identified in three steps. Furthermore, a case-base is established by considering CBR-based model. In the next stage, the CBR-based model is developed in seven steps for cost estimation of drilling wells. Methods of calculating feature weight, feature distinction, case distinction, and adaptation can be different according to viewpoint of researcher's who is developing the CBR model.

3.1. Step 1: Environment of case-base

3.1.1. Step 1-1: Data collection

Data collection is the most important and time consuming stage in projects of data mining. Since the data are the input of projects, the more accurate the input, the more accurate the work output. There are final reports at the end of project for each project which is defined. In this study which we are going to use past data, paying attention to the issue discussed above is important and the art of a data analyst is that he or she be able to extract the best and appropriate features according to the available data.

3.1.2. Step 1-2: Filtering data

A data set consists of data objects. Data set is usually a file in which objects (cases) are the file rows and each column corresponds to one feature. Four types of features can be defined:

- **Nominal**: Names are different merely and only provide the information to distinguish case.
- **Rank**: It provides the enough information to sort the cases.
- **Interval**: Difference between values is meaningful. That is, there is the unit of measurement.
- **Ratio**: Differences and ratios are meaningful.

Nominal and rank features are known as class or qualitative feature all together. These features have some limited modes. Even if these features are expressed with a number for example an integer number, they should be treated as a symbol. Two other types (i.e. interval and ratio) are known as a numerical or quantitative feature. Quantitative features are expressed with numbers and they hold the most of number's properties. These features can adopt the integer or continuous value.

3.1.3. Step 1-3: Normalizing data

Scale of features measurement in data may be different, as an example a feature like length of drilling well contains a greater range than inclination of well. With regard to the difference amount aggregated in features of distance function, high-scale features eliminate the effect of low-scale features. To solve this problem, values should be normalized before comparison. Normalization with regard to the type of feature is explained in step 2-2.

3.2. Step 2: The CBR model

This step shows the CBR-based model for inferring feature weight, features distinction, cases distinction, selecting proposed solutions, adaptation, and model validation.

3.2.1. Step 2-1: Calculating feature weight

In this study, we use the method of Montazemi and Gupta [5] to calculate the weight of feature.

$$W_j = \frac{\sigma_j}{\sum_I \sigma_j} \; I \in A^k \cap A^i, \forall j. \; [1], \; \sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(S_{ik}^j - \bar{S}^j)^2} \; \forall j, \; [2], \; \bar{S}^j = \frac{1}{m}\sum_{i=1}^{m} S_{ik}^j \tag{1}$$

In which A^k and A^i are, respectively, set of the features of new case and previous cases (case-base) and S_{ik}^j is similarity of feature between new case k and case i in case-base. Since we use d_{ik}^j in our formulas (steps 2-2 and 2-3). With placement $S_{ik}^j = 1 - d_{ik}^j$ in the above formulas:

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{d}^j - d_{iK}^j)^2} \quad \forall j, \tag{2}$$

3.2.2. Step 2-2: Calculating distinction with regard to the type of feature

In CBR, the purpose of numerical distinctions is to quantify the differences which exist between two or more cases or structures. In this section, the distinction of case is explained according to the type of used feature or variable.

3.2.2.1. Interval and ratio features

For each feature j of case i^{th}, we normalize the feature the following:

$$v'_{ij} = \frac{v_{ij} - M_j}{S_j} \; [5], \; M_j = \frac{1}{m}(v_{1j} + v_{2j} + \ldots + v_{mj}), \; S_j = \frac{1}{m}(|v_{1j} - M_j| + |v_{2j} - M_j| + \ldots + |v_{mj} - M_j|)$$

In which v_{ij} is the feature value j in case i^{th}. Now, we can measure the distinction between cases with one of methods of distances measurement. One of famous distances is Minkowski distance:

$$d_{iz}^{\text{Interval\&Ratio}} = (|v'_{i1} - v'_{z1}|^q + |v'_{i2} - v'_{z2}|^q + \ldots + |v'_{ir} - v'_{zr}|^q)^{\frac{1}{q}} \tag{6}$$

In which $r (r \le n)$ is the number of features of the type of interval and ratio in two cases i and z. For $q = 1$ and $q = 2$, respectively, Manhattan and Euclidean distance is obtained. If we define W_j for each feature:

$$d_{iz}^{\text{Interval\&Ratio-Weighted}} = (W_1 |v'_{i1} - v'_{z1}|^q + W_2 |v'_{i2} - v'_{z2}|^q + \ldots + W_r |v'_{ir} - v'_{zr}|^q)^{\frac{1}{q}} \tag{7}$$

3.2.2.2. Nominal features

Nominal features are divided to two modes: 1) The Combo Box mode: we can choose only one mode and value from the total number of modes of the considered features. 2) The List Box mode: we are allowed to choose several modes and value from the total number of modes of the considered features.

In the Combo Box mode, if s is the number of modes of the nominal feature, then the position of these modes can be demonstrated with $1, 2, \ldots, s$ numbers. For measuring the distinction between cases with regard to nominal features in this mode, we use the following equation:

$$d_{iz}^{Nominal(C)} = \frac{t - s}{t} \tag{8}$$

In which s is the number of features that cases i and z have the same modes from that feature and $t\,(t \le n)$ is the total number of nominal features in the Combo Box mode. If we define w_j for each feature:

$$d_{iz}^{Nominal(C)-\text{Weighted}} = \frac{t-s}{t}\left(W_1+W_2+...+W_t\right) \tag{9}$$

And in the List Box mode, if s_j is the number of modes of the nominal feature j^{th}, then the position of these modes can be demonstrated with $1,2,...,s_j$ numbers. To measure the distinction between cases with regard to nominal features in this mode, we use the following equation:

$$d_{iz}^{\text{Nominal(L)}} = \frac{\sum_{j=1}^{p} Max(t_j^i, t_j^z) - \sum_{j=1}^{p} s_j}{\sum_{j=1}^{p} Max(t_j^i, t_j^z)} \tag{10}$$

In which $p\,(p \le n)$ is the number of nominal features, s_j is the number of the same modes of feature j^{th} between cases i and z, t_j^i and t_j^z, are, respectively, the number of obtained modes (not the number of total modes (t_j)) of nominal feature j^{th} in cases i and z. If we define w_j for each feature:

$$d_{iz}^{Nominal(L)-\text{Weighted}} = \frac{\sum_{j=1}^{p} Max(t_j^i, t_j^z) - \sum_{j=1}^{p} s_j}{\sum_{j=1}^{p} Max(t_j^i, t_j^z)}\left(W_1+W_2+...+W_p\right) \tag{11}$$

3.2.2.3. Rank features

In these features, the sequential value of each position is specified but the distance between these positions is meaningless. Suppose that the number of different modes of rank features j is $1,2,...,O_j$. Calculating the distinction of cases based on these features includes three steps:

- Step 1) replace v_{ij} with number of its sorted position in j. i.e., $r_{ij} \in \{1,2,...,O_j\}$ which r_{ij} is a rank that is assigned to v_{ij}.
- Step 2) since rank features have different ranges, so we normalize them to [0, 1] through the following equation:

$$Z_{ij} = \frac{r_{ij}-1}{O_j-1} \tag{12}$$

  That O_j is the maximum of possible modes of rank feature j.
- Step 3) now each of the methods of measuring distance can be used.

$$d_{iz}^{\text{Rank}} = \left(\left|Z_{i1}-Z_{z1}\right|^q + \left|Z_{i2}-Z_{z2}\right|^q + ... + \left|Z_{il}-Z_{zl}\right|^q\right)^{\frac{1}{q}} \tag{13}$$

In which $l\,(l \le n)$ is the number of features of the type of rank in two cases i and z. If we define w_j for each feature:

$$d_{iz}^{\text{Rank-Weighted}} = \left(W_1\left|Z_{i1}-Z_{z1}\right|^q + W_2\left|Z_{i2}-Z_{z2}\right|^q + ... + W_l\left|Z_{il}-Z_{zl}\right|^q\right)^{\frac{1}{q}} \tag{14}$$

3.2.2.4. Symmetric and Asymmetric Binary features

Binary features (0 and 1) are two types, symmetric and asymmetric. Unlike symmetric features in asymmetric features, only existence (or non-zero value) is important. To measure the distinction between cases i and z, we form the distinction matrix of Fig. 2 for both symmetric and asymmetric.

Case i \ Case z	1	0
1	a	b
0	c	d

Fig. 2: distinction matrix

In which $g = a+b+c+d$ is the number of features of the type of symmetric or asymmetric and a is the number of features that their values in both case i and z is equal to 1 and similarly, b, c and d are

defined according to Fig. 2. To calculate the distinction between case i and z , if all of binary features are symmetric, we use the following equation:

$$d_{iz}^{\text{Symmetric Binary}} = \frac{b+c}{a+b+c+d} \tag{15}$$

If we define w_j for each feature:

$$d_{iz}^{\text{Symmetric Binary-Weighted}} = \frac{b+c}{a+b+c+d}\left(W_1+W_2+...+W_g\right) \tag{16}$$

To calculate the distinction between case i and z for asymmetric binary features, we use the following equation:

$$d_{iz}^{\text{Asymmetric Binary}} = \frac{b+c}{a+b+c} \tag{17}$$

(Note that d has omitted because the negative features or with zero value for both cases i and z have the little importance according to the contract). If we define w_j for each feature:

$$d_{iz}^{\text{Asymmetric Binary-Weighted}} = \frac{b+c}{a+b+c}\left(W_1+W_2+...+W_{g-d}\right) \tag{18}$$

3.2.3. Step 2-3: Calculating cases distinction

If cases have the various features, the distinction of features summation is their average individual distinctions. The distinction between cases i and z is defined as following equation:

$$d(i,z) = \frac{d_{iz}^{\text{Interval\&Ratio}} + d_{iz}^{Nominal(C)} + d_{iz}^{Nominal(L)} + d_{iz}^{\text{Rank}} + d_{iz}^{\text{Symmetric Binary}} + d_{iz}^{\text{Asymmetric Binary}}}{NOF} \tag{19}$$

In which $1 \le NOF \le 6$ is the number of the type of feature in our model. If we define w_j for each feature:

$$d(i,z) = \frac{d_{iz}^{\text{Interval\&Ratio-Weighted}} + d_{iz}^{Nominal(C)-\text{Weighted}} + d_{iz}^{Nominal(L)-\text{Weighted}} + d_{iz}^{\text{Rank-Weighted}} + d_{iz}^{\text{Symmetric Binary-Weighted}} + d_{iz}^{\text{Asymmetric Binary-Weighted}}}{\sum_{j=1}^{n} W_j} \tag{20}$$

3.2.4. Step 2-4: Selecting proposed solutions

In step 2-4, the cases are ranked from the lowest to the highest according to their percentage of distinction which are created in step 2-3. The case with the lowest distinction is most similar to new case. To use adaptation, we need several similar cases. So the cases which their percentage of distinction is not more than a percentage which is important for us are considered as proposed solutions.

3.2.5. Step 2-5: Adaptation

Adaptation looks for outstanding differences between retrieved case and new case, then applies formulas or rules that take these differences into account when suggesting a final solution. In here, we use the automatic adaptation method in most stages because the CBR system which we write its program in C#, can adapt the retrieved solutions. And at the end, we use the manual adaptation method to avoid features value which may change the nature of new case. Adaptation method is as following:

We put the most similar case (with the lowest percentage of distinction) as criterion which in step 2-4 is obtained. Manager believes that the percentage of distinction of the most similar case with new case should not exceed % β. Thus, we should try to reduce the percentage of distinction to% β (if it is more than % β) by use of adaptation method. In our proposed model, pseudocode of adaptation method is as following:

In next for, j is index of the features; n is number of the features; for (j=1; j< n; j++) in next if, ms is index of the most similar case; k is index of the new case; {if ($d_{ms,k}^{j}$ > β%) minimum d index= ms ; in next for, i is index of the cases; p is index of proposed solutions; {for (i=1;i< p; i++) {if ($d_{ik}^{j} < d_{ms,k}^{j}$) { $d_{ms,k}^{j} = d_{ik}^{j}$; minimum d index= i ;}}}} the $d(ms,k)$ is calculated; in next for, j is, respectively, the highest weight; e is number of examined features at the bove; for (j=1; j< n-e; j++) {if ($d(ms,k)$ > β%) minimum d index= ms ; {for (i=1;i< p; i++) {if ($d_{ik}^{j} < d_{ms,k}^{j}$) { $d_{ms,k}^{j} = d_{ik}^{j}$; minimum d index= i ;}}} the $d(ms,k)$ is calculated; $d(ms,k) = d(ms,k)$; else {is obtained final solution;}} in next for, j is, respectively, the highest weight; for (j=1; j< n; j++) {if

($d(ms,k) > \beta\%$) minimum d index= ms ; in next for, m is number of the cases except proposed solutions;{for (i=1;i< m; i++) {if ($d^j_{ik} < d^j_{ms,k}$) { $d^j_{ms,k} = d^j_{ik}$; minimum d index= i ;}}} the $d(ms,k)$ is calculated; $d(ms,k) = d(ms,k)$; else {the final solution is obtained;}}.

At the end, we use manual adaptation method as follows: we compare each of feature of new modified case with its corresponding feature in new case and if feature value of the new modified case naturally conflicts with its corresponding feature value in new case, we'll modify its value with regard to experience and knowledge which we have about the whole system and consequently from the whole case.

3.2.6. Step 2-6: Obtaining output feature of final solution

In this step, we obtain the estimated output or cost of final solution as following:

In written programming, it is specified that value of each features of final solution obtained in Step 2-5 is associated with which case. For example, if our output is the cost, then we extract the cost of each of features from relevant case (which most of them are from the most similar case to new case). Of course, it is noteworthy that we should calculate these costs in the present with regard to time and date of implementing and cost present value of it feature. At the end, we sum all costs. Consequently, the estimated cost of new case is obtained.

3.2.7. Step 2-7: Model validation

Step 2-7 calculates the estimation accuracy. This section compares the new case with final solution (final case), which is obtained after the adaptation stage, and calculate the standard error rate (SER) and the estimation accuracy (EA). Equation 21 is used for calculating the SER.

$$SER = \frac{\left| C_{Actual-NewCase} - C_{CBR-Final\ Solution} \right|}{C_{Actual-NewCase}} \times 100\,\% \tag{21}$$

In which $C_{Actual-New\ Case}$ is the actual cost of new case, and $C_{CBR-Final\ Solution}$ is the cost of final solution using CBR that is obtained in step 2-6. Equation 22 is used for calculating the EA:

$$EA = 100 - SER \tag{22}$$

4. Case study

At the present study, we show the application of CBR method for cost estimation of the drilling of oil and gas wells in design phase in Petropars Company. A case-base is made with 10 cases (wells) and is shown in Table 1. Features of drilling project of wells are inferred through extensive literature review, and more importantly, based on final reports of different parts and important and variable effect on which they have in different wells cost (the parts which almost are used in all drilling projects both onshore and offshore and or have fixed cost, are not considered as feature) and also interview with experts in the field drilling. The names and type of the scale of each feature related to cases of drilling wells are shown in Table 1.

Table 1: the case-base

features	Scale	Well 1	…	Well 10	New Well
Well Type	Nominal (C)	appraisal	…	development	development
Exploratory Material	Nominal (C)	Gaz	…	Gaz	Gaz
Rig Drow Works (horse power)	Ratio	3000	…	3000	3000
Drilling Length (m)	Ratio	4100	…	4756.8	4515
Operation Time (hours	Ratio	1900	…	2037	1950
Non Productive Time (hours)	Ratio	130	…	137	150
Average GPM Interval	Ratio	850	…	1000	965
Average Pressure Interval	Ratio	1150	…	1200	1350
Rig Rotary	Ratio	45	…	42.5	43.5

System (Average RPM Interval)					
Rig Mud Pump Type	Nominal (C)	douplex	…	douplex	douplex
Volume of Rig's Mud Tanks	Ratio	900	…	800	700
Rig Well Control System (BOP)	Rank	2000or3000(psi)	…	5000 psi	5000 psi
Logging	Nominal (L)	LWD, MWD	…	LWD, MWD	LWD, TLC
Cementing additive (Cement Class)	Nominal (L)	B, D, G	…	A, B, C	C, E, G
Consumed Cement (MT per Bulk)	Ratio	400	…	366	395
Stimulation Services	Nominal (L)	acid wash	…	acid wash, matrix acidizing	acid wash, matrix acidizing
Coring	Asymmetric Binary	YES	…	NO	NO
Perforation	Nominal (L)	CIRP, CTTCP	…	openhole with 7"Liner, TCP	CTTCP, SPF
Well Profile	Nominal (L)	deviated	…	horizontal, multi lateral	deviated
Casing / Tubing Running	Rank	18_5/8",13_3/8",10_3/4",9_5/8",26"CP,7"Liner	…	18_5/8",13_3/8",10_3/4",9_5/8",26"CP,7"Liner	18_5/8",13_3/8",10_3/4",9_5/8",26"CP, 7"Liner,2_7/8"-3_1/2"TBG
Fishing Services	Asymmetric Binary	YES	…	YES	YES
Slick Line	Asymmetric Binary	YES	…	YES	YES
Downhole Completion Equipment	Nominal (C)	monobor	…	monobor	monobor
Bits	Nominal (L)	rock, Milled Tooth	…	rock, Insert	button, milled Tooth
MUD	Nominal (L)	oil base,water base	…	oil base	gas base(compressed air)
MLS-Services	Asymmetric Binary	YES	…	NO	NO
Percent CO2	Ratio	2	…	2.4	2
Percent H2S	Ratio	0.5	…	0.3	0.5
Producing Reservoir	Ratio	K-1, K-2, K-3, K-4	…	K-1, K-2, K-3	K-1, K-2, K-3, K-4
Water Depth	Ratio	67.8	…	75	60
Average Inclination (α)	Interval	57	…	49	49
Average Azimuth (α)	Interval	240	…	254.7	253.68
Max Reservoir Pressure (psi)	Ratio	4800	…	5200	4960
Well Cost(million dollars)		50	…	43	**?**

In next stage, we implement the CBR-based model for cost estimation of drilling wells. For this purpose, we initially calculated the distinction with regard to type of feature. Is obtained the features weight. Then we calculated the cases distinction. Case No. 3 is the most similar case (with distinction percentage 29.5%). To use adaptation, we need several similar cases. So cases are considered as proposed solutions that their distinction percentage is not more than 35% (a percentage which is important for us) that hereby, cases No. 1, 3, and 10 are retrieved. Manager believes that the percentage of distinction for final solution should not be more than % 13. The CBR system which we write its program in C#, adapted the retrieved solutions. The final solution did not need the manual adaptation because there was no feature value that could change the nature of new case. The percentage of distinction is reduced to 11.1%. We obtain the estimated drilling cost of new well or cost of final solution as follows:

We extracted the cost of each feature of final solution from relevant case (which most of them are from case No. 3). These costs are calculated at present with regard to time and date of implementing and cost

present. At the end, we summated all costs. Consequently, the estimated cost of new well is obtained 45.6 million dollars.

Now in order to model validation, SER and EA are calculated using equations 21 and 22. For example, SER and EA are in retrieved case No. 1:

$$SER = \frac{|50-46.2|}{50} \times 100 = 7.6\% \quad and \quad EA = 100 - 7.6 = 92.4\%$$

5. Conclusion

In this study, we estimated the cost of drilling wells using CBR method. Cost estimation accuracy which was an important factor in success of drilling project, is obtained 92.4% for a test case in case base. Therefore our CBR model accuracy is high and the model is useful. Unlike other methods, we did not need output data for all defined cases in case base, which in this study output data is the cost of drilling wells, because our purpose is primarily to obtain the cases distinction which for this purpose, we did not need the wells cost. Initially we needed only the cost of the most similar case. Then in stage of adaptation, the cases are specified which we need their costs. In near future, fuzzy similarity combined with CBR will be considered as interesting topics in costs estimation.

References

[1] S. Perera, and I. Watson. Collaborative case-based estimating and design. Advances in Engineering. 1998, 29 (10): 801–810.

[2] P. Duverlie, and J.M. Castelain. Cost Estimation During Design Step: Parametric Method versus Case Based Reasoning Method. Int J Adv Manuf Technol. 1999, 15: 895–906.

[3] G.H. Kim, S.H. An, K.I. Kang. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and Environment. 2004, 39: 1235 – 1242.

[4] P.C. Chang, J.J. Lin, W.Y. Dzan. Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models. J Intell Manuf, 2010.

[5] A.R. Montazemi, and K.M. Gupta. A framework for retrieval in case-based reasoning systems. Annals of Operations Research. 1997, 72: 51 – 73.