DATA MINING-BASED PREDICTIVE MODEL TO DETERMINE PROJECT FINANCIAL SUCCESS USING PROJECT DEFINITION PARAMETERS

Seungtaek Lee, Changmin Kim, Yoora Park, Hyojoo Son, and Changwan Kim*

Department of Architecture Engineering, Chungang University, Seoul, Korea
** Corresponding author (changwan@cau.ac.kr)*

ABSTRACT: The planning stage is important for project development because the majority of important decisions are made at this stage. Having a well-defined project plan will reduce project uncertainty and increase the likelihood of the project's success. In other words, based on the level of project definition in the planning stage, project success or failure can be predicted. The aim of this study is to generate a predictive model that will forecast project performance in terms of cost, depending on the project definition level during the early stages of the project before a detailed design is started. The predictive model for this study was generated by support vector machine (SVM). A survey of 77 completed construction projects in Korea was conducted in order to collect the project defined level and cost data from each of those projects by questioning selected clients, architects, and construction managers who had participated before beginning the detailed design stage in the project. It is anticipated that prediction results will help clients and project managers revise their project planning when they encounter a poor performance prediction. Furthermore, the research result imply that employing the proposed model can help project participants achieve success by managing projects more effectively.

***Keywords**: Cost Performance, Data Mining, Performance Prediction, Project Definition Rating Index, Project Planning, Support Vector Machine*

1. INTRODUCTION

The construction industry is characterized by high levels of risks and uncertainties. Over the past several decades, many construction projects have experienced large variations in cost and/or schedule [1]. To prevent cost overruns and schedule delays, it is important to have an early understanding of the likelihood of the project's success [2]. It is implicitly assumed that when a contractor or project manager can predict the amount of cost overruns, the prediction of project performance will help project participants to make important decisions [3].

Several research studies have investigated the issue of performance prediction of construction projects. Dissanyaka and Kumaraswamy [4] predicted project performance by using project characteristics, procurement system, project team performance, contractor characteristics, design team characteristics, and external conditions. Kim et al. [2] predicted project success of international construction projects using project condition, ability of the owner and A/E, the contractor's capability and experience, quality of design, and capability of claim. However, it is more appropriate to use project definition elements to predict project performance in planning stage rather than the factors used in the aforementioned research studies because project scope definition is a key element in the pre-project planning process and known to simultaneously correlate to the achievement of excellent project performance [5,6].

Project scope definition is the process included in the pre-project planning process by which projects are defined and prepared for construction [7]. If the project definition is not sufficiently prepared in the pre-project planning process, unexpected changes may occur, rework is required and project rhythm is interrupted. In other words, the project productivity and the morale of the worker may decrease. Thus, the project may earn lower profits or even incur a loss due to an unclear definition of the scope of work [6,8]. In other words, as the project definition directly affects project performance, the success of the project highly depends on the degree of project definition. Wang and Gibson [6] predict project performance using the Project Definition Rating Index (PDRI) score by employing an artificial neural network. Using PDRI scores to predict project performance has merits that can predict performance easily, but it is hard to identify which definition elements affect the performance and address the problem of taking corrective actions in order to revise the project. Moreover, according to other performance prediction researches, several kinds of data mining methods were employed including support vector machine, decision tree, and k-nearest neighbor etc. Among the data mining methods, many researches show that SVM outperforms other methods [9,10,11]. Moreover, if less collected data is available, SVM can be a better alternative to build the prediction model [12]. Thus, the objective of this research is to predict cost performance using project definition elements by employing support vector machine method before a detailed design is begun.

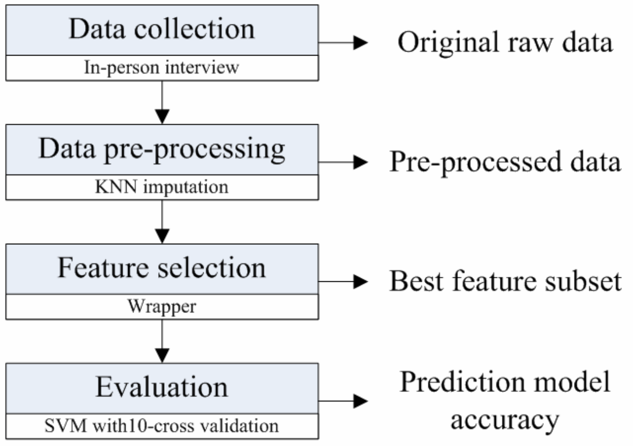

Fig. 1 The framework of the methodology

2. METHODOLOGY

2.1 DATA COLLECTION

Data were collected from 77 completed construction projects in Korea within no more than 3 years of the completion. In order to obtain more reliable and accurate data, an in-person interview was conducted using a questionnaire with each project participant including clients, architects, and projects managers with over ten years of experience in the construction field who had participated in the project planning.

Two sections of the questionnaire were used to measure how the project performance was influenced by various project definition elements. The first part contained questions about project cost and duration in order to measure project performance. Project performance is defined as the extent of variation between the planned and the actual estimates. In this research, cost variation was measured to evaluate project performance. Cost variance was calculated by the following equation:

$$\frac{\text{Actual cost} - \text{ Planned cost}}{\text{Planned cost}} \times 100(\%) \qquad (1)$$

The second section measured the degree of project definition elements on a five-point Likert scale. In order to collect project definition information on project planning, the PDRI was used. The PDRI developed by Construction Industry Institute (CII) can measure the levels of 64 scope definitions to evaluate the project status before detailed design [5,7].

Although data were collected through in-person interviews, there were a few missing values because certain elements were impossible to measure in certain projects. In order to replace missing values, a K-Nearest Neighbor (KNN) imputation method was applied. Due to its simplicity, ease of understanding, and relatively high accuracy, the KNN imputation has been widely used in diverse real applications. The missing value is replaced with the one most frequently found among the *k* number of the most similar data [13]. Among the total data, 3.25% missing value was replaced by KNN imputation method.

2.2 FEATURE SELECTION

The main objective of the feature selection method is to remove redundant features and to select relevant subsets of features to improve prediction accuracy. In this research, the wrapper feature selection method was applied. The wrapper method generates optimal candidate feature subsets, and evaluates through a predetermined data mining method [14]. This method searched for the most appropriate subset of features to each data mining method; more accurate prediction results to each data mining method were expected. Thus, for this paper, the wrapper method has been adapted as a select feature subset.

2.3 SVM PREDICTION MODEL

2.3.1 THE PRINCIPLE OF SVM

The support vector machine (SVM) has recently been a well-used method for data mining in order to apply classification and regression problems. The support vector machine, developed by Vapnik [15], transforms the data into a high dimensional feature space by using kernel mapping in order to better explain the relationship between input and output variables [16]. SVM was originally developed for classification and was later designed to solve regression problems, using Support Vector Regression (SVR). Thus, using SVR, regression problems can be solved by the support vector machine. In SVR, the original regression model approximates the function using the following form:

$$f(x) = (\omega, x) + b \qquad (2)$$

The Euclidean norm (i.e., $\|\omega\|^2$) must be minimized. Formally, this can be written as a convex optimization problem by requiring:

$$\text{minimize} \quad \frac{1}{2}\|\omega\|^2 \qquad (3)$$

$$\text{subject to} \quad \begin{cases} y_i - (\omega, x_i) - b \le \varepsilon \\ (\omega, x_i) + b - y_i \le \varepsilon \end{cases}$$

The convex optimization problem is feasible in cases where *f* actually exists and approximates all pairs (x_i, y_i) with ε precision. However, in some cases, errors outside the margin of ε-tubes are allowed. Introducing slack ξ_i, ξ_i^* to cope with otherwise infeasible constraints of the optimization problem Equ (6), the formulation becomes

$$\text{minimize} \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \qquad (4)$$

$$\text{subject to} \quad \begin{cases} y_i - (\omega, x_i) - b \le \varepsilon + \xi_i \\ (\omega, x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

ε-insensitive loss function has been described by

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \textit{otherwise} \end{cases} \qquad (5)$$

Therefore, Equ. (4) can be written in the following explicit form, Equ. (6):

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \qquad (6)$$

Any function that satisfies Mercer's condition can be employed as a kernel function $k(x_i, x)$ [15]. Employing some diverse Kernel functions (polynomial, radial basis function, two-layer neural network, etc.) can solve not only linear relationships but non-linear relationships as well.

2.3.2 MODELING OF SVM

In case of the SVM model, including complexity parameter C and RBF kernel parameter γ has an influence on the performance of the technique. C determines the trade-off between the empirical risk and the regularization term, which is the ability of prediction for the technique. If C is too large, the model will focus on reducing the empirical risk instead of the model capacity [17]. The RBF kernel parameter (γ) can affect the decision boundary shape, thus, it influence the generalization ability of the SVM [18]. Thus, when building an SVM prediction model, optimum parameters should be found.

In this research, 10-cross validation was used to evaluate SVM prediction accuracy. This method divides total data set into 10 data sets. Among the 10 divided data sets, nine data sets are used for the build prediction model and the remaining data set is used as a test set to evaluate performance of the model. The accuracy of the performance is calculated after 10 repetitions of the process. Thus, it can be expected that a reliable result is obtained [19].

3. EXPERIMENTAL RESULT

3.1 PERFORMANCE CRITERIA

The overall performance of data mining methods was estimated in terms of correlation coefficient (R), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). These values are measured to evaluate error between predicted and actual values. The R, MAE, and RMSE are defined as:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}} \qquad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \qquad (8)$$

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (9)$$

Notes. x_i = actual project cost performance value (%), y_i = predicted project cost performance value (%), $\bar{x}_i$ = mean of actual project cost performance value (%), n=the number of data

The correlation coefficient is shown as a linear relation between the actual and predicted cost performance. If the correlation coefficient value is closer to either 1 or −1, it means that there is a strong relation between the values. The ideal value of RMSE and MAE is zero when predicted values and actual values are exactly same. The difference between the MAE and the RMSE is whether or not they are influenced by outliers. The RMSE is affected by outliers; it generally exceeds MAE to the extent of the outliers' value. Thus, if there are many big outliers, RMSE values become much higher than MAE.

3.2 RESULTS OF SVM

In this paper, feature selection was initially applied and a support vector machine was used with the selected variables. Out of the total 64 project definition elements, the wrapper feature selection was applied to find the relevant project definition elements. As a result, 39 optimal features were selected

Since the accuracy of the SVM prediction model is largely dependent on the selection of the parameters, parameter C and RBF kernel parameter γ, it is important to optimize the parameters. Because it cannot know which value of parameter is the optimum for the prediction model, in this research, a grid search was conducted to search for the best pair of C and γ values. The parameter pair which showed the lowest mean absolute error was selected. As a result, the best parameter pair was found as $C=2^{12}$ and $\gamma = 2^{-12}$ which gives the lowest mean absolute error of 4.72, Moreover, the parameter also gives the lowest root mean's squared error of 6.61 and the highest correlation coefficient of 0.8. Figure 2 represents the variation of actual and predicted cost performance.

3.3 COMPARISON OF SVM AND OTHER PREDICTION MODELS

In this research, Multiple Linear Regressions (MLR), K-

Nearest Neighbor (KNN), Decision Tree (DT), and Artificial Neural Network (ANN) were also applied to compare performance of the support vector machine. When the prediction methods were applied, the methods also used only the relevant feature subset selected by the wrapper method. Moreover, the parameter of each prediction method also significantly affected the outcome of the prediction method. Table 1 provides the optimal values of parameters for this data set with KNN, DT, and ANN. Through the optimal parameter values, the prediction models are built and evaluate the accuracy of the performance. Table 2 compares the performance of the five prediction models using 10-cross validation. As shown, SVM outperforms other prediction methods which show the lowest MAE and RMSE of 4.72 and 6.61, respectively, and the highest correlation coefficient of 0.8, which is considered a strong correlation. This demonstrates that the SVM method is a better cost prediction model than others.

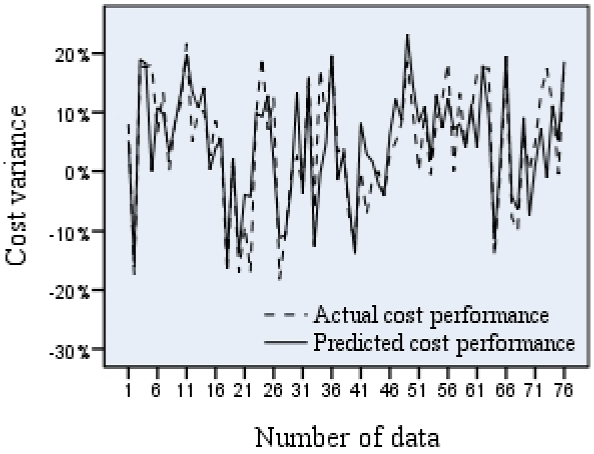

Fig. 2 Variation in actual and predicted cost performance

Table 1. Optimal value of parameters of the data mining methods

Method	Parameters	Values
KNN	Number of neighbors	10
DT	Min number of instance	7
ANN	Nodes in the hidden layer	4
	Learning rate	0.13
	Momentum	0.003
	Training epochs	1000

Table 2. Comparison accuracy between other models

Method	R	MAE	RMSE
SVM	**0.80**	**4.72**	**6.61**
MLR	0.62	6.93	8.60
KNN	0.78	5.31	6.86
DT	0.51	7.27	9.38
ANN	0.61	6.83	8.59

4. CONCLUSION

This paper presents a support vector machine to predict project cost performance before the detailed design stage. The result shows that the SVM prediction model serves as a realistic model that gives low MAE and RMSE of 4.72 and 6.61, respectively, and a high correlation coefficient of 0.8, which signifies a strong correlation. Thus, the prediction model can provide project participants with a helpful and useful guide to the likely cost overrun for the project. Based on the results, project participants can better manage the project to improve cost performance. Moreover, project participants can also pay closer attention to the selected project definition elements. Thus, the project planning stage can be utilized more efficiently, and the project is more likely to finish successfully. In this research, although the model with data from 77 projects shows appropriate results, if there are more data, more accurate prediction result can be expected. In addition, further research can be conducted to predict other project performance (i.e., project schedule, owner satisfaction, etc.) using the project definition elements.

ACKNOWLEDGEMENT

This research is supported by a grant (07UrbanRenaissanceB03) from the High-Tech Urban Development Program funded by the Korean Ministry of Land & Transport and Maritime Affairs.

REFERENCES

[1] Abdelgawad, M. and Fayek, A.R., "Risk Management in the Construction Industry Using Combined Fuzzy FMEA and Fuzzy AHP", *Journal of Construction Engineering and Management*, Vol. 136(9), pp. 1028–1036, 2010.

[2] Kim, D.Y., Han, S.H., Kim, H., and Park, H., "Structuring the Prediction Model of Project Performance for International Construction Projects: A Comparative Analysis", *Expert Systems with Applications*, Vol. 36(2), pp. 1961–1971, 2009.

[3] Ling, F.Y.Y., Chan, S.L., Chong, E., and Ee, L.P., "Predicting Performance of Design-Build and Design-Bid-Build Projects", *Journal of Construction Engineering and Management*, Vol. 130(1), pp. 75–83, 2004.

[4] Dissanayaka, S.M. and Kumaraswamy, M.M., "Comparing Contributors to Time and Cost Performance in Building Projects", *Building and Environment*, Vol. 34(1), pp. 31–42, 1999.

[5] Construction Industry Institute (CII), *Pre-Project Planning Tool: PDRI for Buildings*, Research Summary, 155–1, Austin, TX, 1999.

[6] Wang, Y.R. and Gibson, G.E., "A Study of Preproject Planning and Project Success Using ANNs and Regression Models", *Automation in Construction*, Vol. 19(3), pp. 341–346, 2010.

[7] Cho, C. and Gibson, G.E., "Building Project Scope Definition Using Project Definition Rating Index", *Journal of Architectural Engineering*, Vol. 7(4), pp. 115–125, 2001.

[8] O'Connor, J.T. and Vickroy, C.G., *Control of Construction Project Scope*, Source Document 6, Construction Industry Institute (CII), Austin, TX, 1986.

[9] Li, Q., Meng, Q., Cai, J., Yoshino, H., and Mochida, A., "Applying Support Vector Machine to Predict Hourly Cooling Load in the Building", *Applied Energy*, Vol. 86(10), pp. 2249–2256, 2009.

[10] Lee, Y.C., "Application of Support Vector Machines to Corporate Credit Rating Prediction", Expert Systems with Applications, Vol. 33(1), pp. 67–74, 2007.

[11] Min, J.H. and Lee, Y.C., "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernal Function Parameters", Expert Systems with Applications, Vol. 28(4), pp. 603–614, 2005.

[12] Behzad, M., Asghari, K., and Coppola, E.A., "Comparative Study of SVMs and ANNs in Aquifer Water Level Prediction", Journal of Computing in Civil Engineering, Vol. 24(5), pp. 408–413, 2010.

[13] Hulse, J.V. and Khoshgoftaar, T.M., "A Comprehensive Empirical Evaluation of Missing Value Imputation in Noisy Software Measurement Data", The Journal of System and Software, Vol. 81(5), pp. 691–708, 2008.

[14] Kohavi, R. and John, G.H., "Wrappers for Feature Subset Selection", Artificial Intelligence, Vol. 97(1–2), pp. 273–324, 1997.

[15] Vapnik V., The Nature of Statistical Learning Theory, Springer, New York, 1995

[16] Ali, S. and Smith, K.A., "On Learning Algorithm Selection for Classification", Applied Soft Computing, Vol. 6(2), pp. 119–138, 2006.

[17] Cherkassky, V. and Ma, Y., "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression", Neural Networks, Vol. 17(1), pp. 113–126, 2004.

[18] Louis, B., Agrawal, V.K., and Khadikar, P.V., "Prediction of Intrinsic Solubility of Generic Drugs Using MLR, ANN and SVM Analyses", European Journal of Medicinal Chemistry, Vol. 45(9), pp. 4018–4025, 2010.

[19] Chu, A., Ahn, H., Halwan, B., Kalmin, B., Artifon, L.A., Barkun, A., Lagoudakis, M.G., and Kumar, A., "A Decision Support System to Facilitate Management of Patients with Acute Gastrointestinal Bleeding", Artificial Intelligence in Medicine, Vol. 42(3), pp. 247–259, 2008.