

Applying Neural Network Ensemble Concepts for Modelling Project Success

Yu-Ren Wang¹, Yi-Jao Chen² and C.F. Jeffrey Huang³

¹Assistant Professor, Department of Civil Engineering, National Kaohsiung University of Applied Sciences, 415 Chine-Kung Road, Kaohsiung Taiwan, 807; PH (886) 73814526-5250; FAX (886) 73831371; email: yrwang@cc.kuas.edu.tw

²Lecturer, Department of Asset Science, Leader University, No. 188, Sec. 5, An-Chung Rd., Tainan 70946, Taiwan, PH(886)62558019; FAX(886)62558528; email: yjchen@mail.leader.edu.tw

³Associate Professor, Department of Civil Engineering, National Kaohsiung University of Applied Sciences, 415 Chine-Kung Road, Kaohsiung Taiwan, 807; PH (886) 73814526-5234; FAX (886) 73831371; email: jeffrey@cc.kuas.edu.tw

Abstract

Researchers have been studying early planning process since the early 1990's and results from these researches suggest that projects with more early planning efforts are more likely to succeed. This study intends to employ neural networks to build credible models linking preproject planning and project success. Preproject planning status as measured by Project Definition Rating Index (PDRI) is set as the independent variable and schedule/cost performance is set as dependent variable. To enhance the performance of the neural networks model, bootstrap aggregation and boosting algorithms are incorporated in the model development process. The results from these two neural network ensemble models are examined. This research finds out that boosting neural network ensemble models produce better results than bootstrap aggregation neural network ensemble models. Results from both models show that project with better early planning can expect better chance of project success.

Keywords: Preproject Planning, Project Success, ANNs Model

1. Introduction

It has long been recognized by the industry practitioners that efforts made early in the project life cycle will have significant impact on the project outcome. Nevertheless, the early planning process varies significantly throughout industry from one organization to another. In light of this, researchers at the Construction Industry Institute (CII) and University of Texas at Austin have set up several research projects focusing on the early planning process since the 1900's. In the early planning stage of the project life cycle, the process of developing sufficient strategic information with which owners can address risk and decide to commit resources to maximize the chance for a successful project is defined as the pre-project planning process (CII 1995). Preproject planning is a major phase of the project life cycle and it begins after a decision is made by the business unit to proceed with a project concept and continues until the detailed design is developed.

Preproject planning process constitutes a comprehensive framework for detailed project planning and includes scope definition. Project scope definition is the process by which projects are selected defined and prepared for definition. It is a key practice necessary for achieving excellent project performance (Morrow and Yarossi 1994) and is a key element in the preproject planning process. How well preproject planning is performed will affect cost and schedule performance, operating characteristics of the facility, as well as the overall financial success of the project (Gibson and Hamilton 1994). Success during the detailed design, construction, and start-up phases of a project highly depends on the level of effort expended during the scope definition phase as well as the integrity of project definition package (Gibson and Dumont 1996). Therefore, it is important to investigate the relationship between preproject planning and project success with real data from the industry.

In order to measure the preproject planning efforts for each construction project, a scope definition tool, Project Definition Rating Index (PDRI) is incorporated in this research to evaluate the completeness of

project scope definition. The Project Definition Rating Index, developed by CII, is a comprehensive, weighted checklist of crucial scope definition elements that have to be addressed in pre-project planning process. It provides the project team a simple and easy-to-use tool to objectively evaluate the current status of a project during preproject planning phase. Since its development, researchers at the University of Texas at Austin and Construction Industry Institute (CII) have been collecting preproject planning information using the PDRI. For the uniqueness of the different sectors in the construction industry, two versions of the PDRI have been developed specifically for the Industrial and Building sectors.

In addition to preproject planning information collected using the PDRI, project performance (cost and schedule) information was also collected through the data collection process. Enhanced Artificial Neural Networks (ANNs) models, using bootstrap aggregation and boosting algorithms, are selected in this research to investigate the relationship between preproject planning and project performance using the sample project data. The ANNs analysis results from the two different models are also compared.

2. Survey Instrument and Data Collection

A scope definition tool, Project Definition Rating Index (PDRI) is used as a survey instrument in these case studies to measure the preproject planning practices in the industry. The PDRI is developed by CII in 1994 as an easy-to-use preproject planning tool to assist owner and contractor companies to better achieve business, operational, and project objectives (CII 1996). The first version is the PDRI for industrial projects, which is a weighted matrix with 70 scope definition elements (issues that need to be addressed in preproject planning) grouped into 15 categories and further grouped into three main sections. Since its development, the Industrial PDRI has been widely used among the CII member companies. In responding to the needs of the building industry, CII chartered a research team and developed the PDRI for Building Projects in 1999 (CII 1999).

The PDRI provides a means for an individual or team to evaluate the status of a construction project during preproject planning with a score corresponding to the project's overall level of definition. The PDRI helps the stakeholders of a project to quickly analyze the scope definition package and to predict factors that may impact project risk specifically with regard to industrial and building projects (Cho 2000).

In the PDRI survey questionnaires, specific questions were intended to obtain historical and "after the fact" project information. The questionnaires included questions regarding project basics (location, type, budget and schedule), operating information, and evaluation using an unweighted PDRI score sheet. Survey participants were asked to think back at a point just prior to construction document (detailed design) development when they filled out the PDRI evaluation score sheet. The total scores were then calculated based on pre-assigned element weights after the questionnaires were returned. Please refer to CII 1996 and CII 1999 for detailed development of PDRI element weights. Due to the unique nature of these two different sectors, industrial and building projects were examined separately throughout this research investigation.

The data collection was accomplished through a series of retrospective case studies. The sample projects used in this study were obtained from three different sources: previous PDRI research, CII Benchmarking and Metrics research, and institutional organizational (which prefers remaining anonymous) PDRI benchmarking research. Data from 62 industrial projects and 78 building projects, representing approximately \$5 billion in total construction cost, were collected and used to conduct an investigation of the early planning practices in the industrial and building industry. Nevertheless, it is important to note that the collected sample from these three sources is based on organization's volunteering projects and not on a random sample of a known population.

3. Data Analysis and Modeling

Information related to preproject planning practice and project performance is collected through the sample project survey. A database is set up for this research project and is used for data storage and further analysis. The preproject planning status for each surveyed project is measured using the PDRI evaluation. Two project performance aspects are of particular concern for this research: cost and schedule performance. Cost performance and schedule performance are measured by cost and schedule growth. In the survey, respondents were asked to provide estimated costs at the start of construction document development as

well as the actual costs after construction completion. Total cost growth measures total project cost growth as a percentage of the initial estimated project cost. Cost performance was measured by project Cost Growth metric obtained as follow:

$$\frac{\text{Final Cost} - \text{Initial Estimated Cost}}{\text{Initial Estimated Cost}} \quad (1)$$

The total project duration used to calculate project schedule growth was measured from the start date of construction documents development to the date of substantial completion in months. The following equation was used for computing project schedule performance, Schedule Growth:

$$\frac{\text{Final Schedule} - \text{Initial Estimated Schedule}}{\text{Initial Estimated Schedule}} \quad (2)$$

When considering cost performance, it is defined in this research that if the project has a cost growth equal to or less than 0, the project is classified as a successful project. In the mean time, when considering schedule performance, a project is defined as a successful project if the schedule growth is equal to or less than 0. It should be noted that schedule and cost performances are investigated separately throughout this research.

Two different algorithms, bootstrap aggregation and boosting, are used to develop Artificial Neural Networks models for this research to investigate the relationship between the preproject planning, as measured by PDRI scores, and project success, as measured by cost/schedule growth.

4. Artificial Neural Networks

The principle of Neural Networks is based on the assumption that a highly interconnected system of simple processing elements can learn complex interrelationships between independent and dependent variables (Elhag 2004). A typical neural network consists of an input layer, an output layer, and one or more hidden layers. These layers are connected by neurons to form a parallel distributed processing system. Each neuron is viewed as a processing element (PE) that receives inputs and generates outputs through an activation function. Each of the connections between the process elements has an associated weight. Figure 1 shows a typical three-layered neural network with an input layer (I), a hidden layer (H), and an output layer (O).

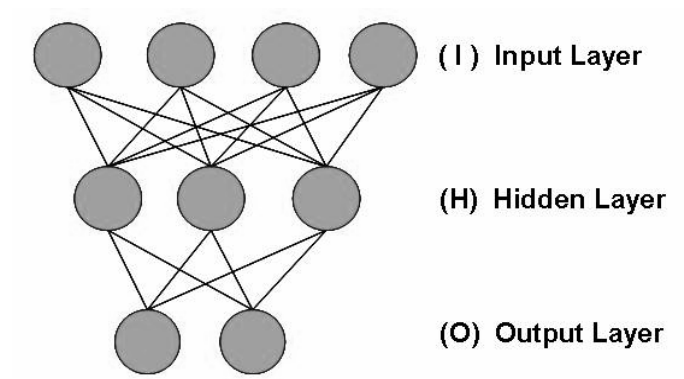


Fig. 1. An example of a three-layered neural network

In the hidden layer, each neuron receives an activation signal (input), and generates a signal (output) through an activation function. The activation signal is the weighted sum of all the signals entering the neuron, as shown in Eq. (3). In Eq. (3), x_j is the activation signal that the neuron j in the hidden layer receives; I_i is i th input in the input layer; and W_{ij} is the weight of the connection between the neuron j in the hidden layer and the input I_i . The neuron (Process Element) produces an output through an activation function that can be any form. The activation function can be either linear or non-linear, and one of the most commonly used activation function is the sigmoid function. The general form of sigmoid function is shown in Eq. (4), where h_j equals output of the neuron j in the hidden layer and x_j equals input for the neuron j .

$$x_j = \sum_i I_i W_{ij} \tag{3}$$

$$h_j = f(x_j) = \frac{1}{1 + e^{-x_j}} \tag{4}$$

$$y_k = \sum_j h_j W_{jk} \tag{5}$$

As presented by Eq. (5), the neurons in the output layer receive activation signals (weighted sum of inputs to neuron k) from the neurons in the hidden layer. In Eq (5), y_k is the input of the neuron k in the output layer and W_{jk} is the weight of the connection between the neurons j and k in the hidden and output layers, respectively. In the output layer, these activation signals are transformed (through activation function) again to generate the outputs of the neural network. This process is shown in Eq. (6), where o_k is the predicted value of the outputs. Then the outputs are compared with desired or actual values, d_k . The error (difference between predicted value and desired/actual value) at the output neurons is defined by Eq. (7). The best performance of the neural network is achieved when the error is minimized.

$$o_k = f(y_k) = \frac{1}{1 + e^{-y_k}} \tag{6}$$

$$E(W) = \frac{1}{2} \sum_k (d_k - o_k)^2 \tag{7}$$

For supervised neural networks (models with specific actual/desired outputs), one of the most effective and popular technique to minimize the error function $E(W)$ is the back-propagation (BP) algorithm. For back-propagation neural networks, the error at the output layer propagates backward to the hidden layer and then to the input layer to update the weights for each of the connections in the neural networks. These forward process (input layer to hidden layer to output layer) and backward process (output layer to hidden layer to input layer) are repeated to minimize the error.

These repeated processes are viewed as learning (training) process. The relationships between inputs and outputs of the system are memorized through the connection weights. It should be noted that before the learning process starts, small random numbers (e.g., between -0.1 and 0.1) are assigned as the initial weights to the connections between the neurons. This ensures that the network is not saturated by large values of the weight, and prevents some training pathologies. Sometimes, the data will be normalized before to obtain convergence within a reasonable number of cycles.

Studies have shown that ANNs have several advantages over the traditional statistical methods such as multiple regression analysis and multivariate analysis (Elhag and Boussabaine 2002). ANNs does not require that the data must follow a specific statistical distribution and does not require predetermination of the relationships between inputs and outputs. In addition, ANNs have very strong capability of self-learning and self-updating. Despite the advantages above, ANNs has been criticized as “unstable algorithm” (Roiger and Geatz 2003) because the model is very sensitive to slight changes in the training data. In view of this, Hansen and Salamon (1990) proposed the concept of neural network ensemble. Instead of developing one neural work to solve a particular problem, a group of neural networks are developed and the network output for each of them are then integrated to obtain a solution for the problem. Bootstrap aggregation and boosting algorithms are two applications of neural network ensembles that can enhance the performance of neural network models. The concept of neural network ensemble (compared with traditional method) is shown in Figure 2 below.

Breiman (1996) used bootstrap method to generate several sub-datasets from the original training data. These sub-datasets are fed into classification tree, composed of a group of classifiers, for training, and the final model output is obtained through the aggregation of the group classifier outputs. It is found bootstrap aggregation is able to reduce the model error rate by 20% to 47% when comparing with single classifier model. For this research, neural networks are created to serve as the classifiers in the bootstrap aggregation (or bagging) method. The outputs from the neural network classifiers are integrated to produce the final output of the model. As illustrated in Figure 3, bootstrap aggregation method is incorporated in the neural network ensemble for this research.

Freud and Schpire (1996) applied boosting algorithms obtained a better result when comparing with

bootstrap aggregation method. Similar to bootstrap aggregation method, boosting method also creates a group of classifiers through a series of training process. However, the selection of training data is different. For bootstrap aggregation method, the training dataset, also known as bootstrap sample, is chosen at random with replacement from the original training set. Boosting maintains a weight for each instance and the higher the weight, the more likely the instance will be selected for classifier training. At each training process, the weight of misclassified instance is increased and thus makes it more likely to be chosen for the next round. After a series of classifier training under boosting approach, the final classifier aggregates the learned classifiers and the results are better than bootstrap aggregation method. Using neural networks as classifiers, the boosting approach is shown in Figure 4.

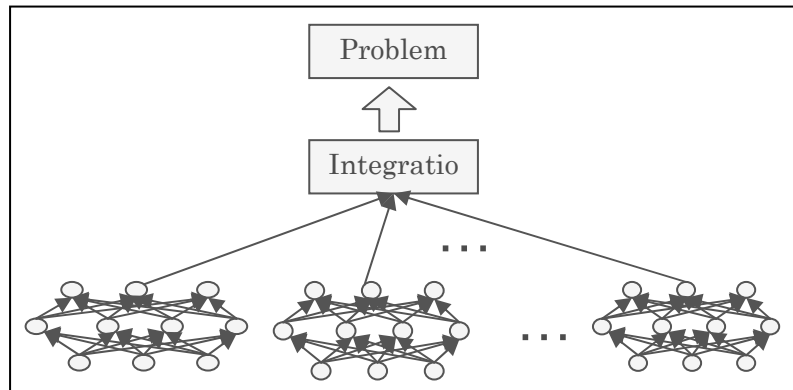


Fig. 2. Neural Network Ensemble

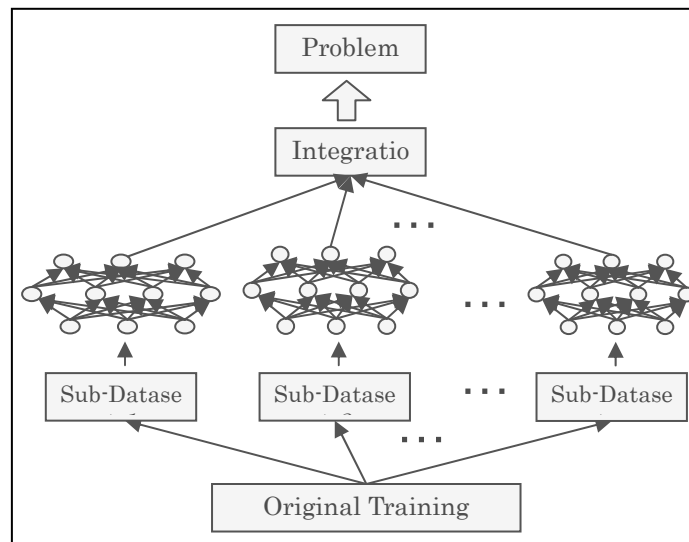


Fig. 3. Bootstrap Aggregation Neural Network Ensemble

Information from a total of 62 industrial and 78 building projects is collected and used to investigate the relationship between the preproject planning and project success. Preproject planning status, as measured by the PDRI score, is the independent variable while the project success, as measured by cost/schedule growth, is the dependent variable in this research. Bootstrap aggregation and boosting algorithms are incorporated to create neural network classifiers using the PDRI scores as model inputs. The sample projects are classified as successful or less-than-successful projects based on their cost and schedule performances.

There are three neural network models developed for this research: single classifier neural network, bootstrap aggregation neural network ensemble model and boosting neural network ensemble model. Single classifier neural network is the traditional neural network model development approach, which uses one neural network classifier for prediction. In this research, the average error rate (misclassification) for single classifier neural network model is 15.79%. Bootstrap aggregation neural network ensemble model trains a group of classifiers and the final model prediction is obtained by voting from the group outputs. The average

error rate is 10.52% under this model. The average error rate obtained for boosting neural network ensemble model is 7.89%.

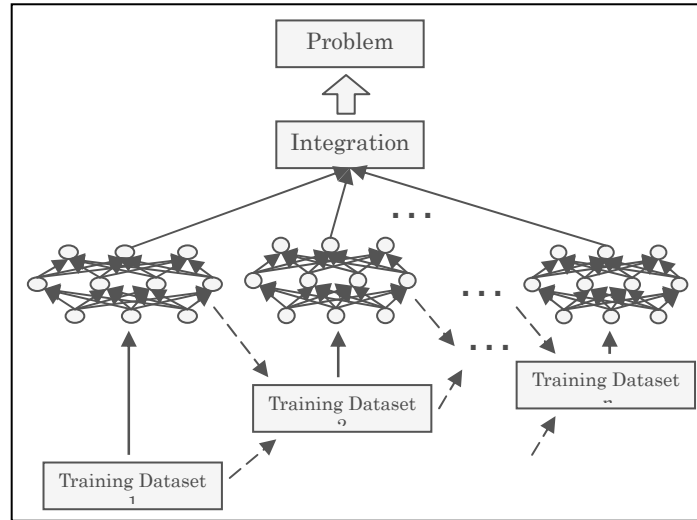


Fig. 4. Boosting Neural Network Ensemble

5. Conclusions

This paper studied the preproject planning of industrial and building construction projects and investigates its relationship with project success (as measured by cost and schedule performance). Questionnaire surveys were used to obtain information related to the status of preproject planning and project performance. Based on the collected data, this research developed three neural network models to classify the projects as successful and less-than-successful projects using the PDRI scores. The first model is traditional single classifier neural network model; the second is bootstrap aggregation neural network ensemble model and the third is boosting neural network ensemble model. The results indicate that boosting neural network ensemble model yields the best classification results. All three models show that projects with better preproject planning (lower PDRI score) are more likely to have a better chance of achieving project success. This paper provides some analytical results for the industry practitioners to put more efforts into the early planning in order to achieve better project performance.

References

- [1] Breiman, L. (1996). "Bagging predictors." *Machine Learning Journal*, 24(2), 123-140
- [2] Cho, C. S. (2000). "Development of the Project Definition Rating Index (PDRI) for Building Projects." Ph.D. Thesis, University of Texas at Austin, Austin, TX
- [3] The Construction Industry Institute (1995). "Pre-Project Planning Handbook". Special Publication 39-2, Austin, TX
- [4] The Construction Industry Institute (1996). "Project Definition Rating Index (PDRI) – Industrial Projects". Implementation Resource 113-2, Austin, TX
- [5] The Construction Industry Institute (1999). "Project Definition Rating Index (PDRI) – Building Projects". Implementation Resource 155-2, Austin, TX
- [6] Elhag, T. M. S., and Boussabaine, A. H. (2002). "Tender price estimation using artificial neural networks. II: Modelling." *J. Finan. Mange. Property Constr.* 7(1), 49–64.
- [7] Elhag, T. M. S. (2004). "Cost modeling: Neural networks vs. regression techniques." *Int. Conf. on Construction Information Technology (INCITE)*, Langkawi, Malaysia, Construction Industry Development Board Malaysia (CIDB).
- [8] Freud, Y. and Schapire R. E. (1996). "Experiments with a New Boosting Algorithm." *Proc. International Conference on Machine Learning*, San Francisco, 148-156
- [9] Gibson, G. E. and Hamilton, M. R. (1994). "Analysis of Pre-Project Planning Effort and Success Variables for Capital Facility Projects", A report to the Construction Industry Institute, Source Document 105, Austin, TX

- [10] Gibson, G. E. and Dumont, P.R. (1996). "Project Definition Rating Index (PDRI)" A report to the Construction Industry Institute, Research Report 113-11, Austin, TX
- [11] Hansen, L. K. and Salamon P. (1990)." Neural Network Ensembles" IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), 993-1001
- [12] Merrow, E. W., and Yarossi, M. E. (1994). "Managing Capital Projects: Where Have We been – Where are We Going?" Chemical Engineering, October: 108-111
- [13] Roiger, R. J. and Geatz, M. W. (2003). "Data mining: a tutorial-based primer." Addison Wesley, Boston.