Automation in Construction 19 (2010) 917–928

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Decision tree for selecting retaining wall systems based on logistic regression analysis

Myungseok Choi, Ghang Lee *

Department of Architectural Engineering, Yonsei University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Accepted 6 June 2010

Keywords:
Retaining wall systems
Logistic regression
Decision tree

ABSTRACT

Machine learning techniques generally require thousands of cases to derive a reliable conclusion, but such a large number of excavation cases are very difficult to acquire in the construction domain. There have been efforts to develop retaining wall selection systems using machine learning techniques but based only on a couple of hundred cases of excavation work. The resultant rules were inconsistent and unreliable. This paper proposes an improved decision tree for selecting retaining wall systems. After retaining wall systems were divided into three components, i.e., the retaining wall, the lateral support, and optional grouting, a series of logistic regression analyses, analysis of variance (ANOVA), and chi-square tests were used to derive the variables and a decision tree for selecting retaining wall systems. The prediction accuracy rates for the retaining walls, lateral supports, and grouting were 82.6%, 80.4%, and 76.9%, respectively. These values were higher than the prediction accuracy rate (58.7%) of the decision tree built by an automated machine learning algorithm, Classification and Regression Trees (CART), with the same data set.

1. Introduction

As a result of rapid urbanization, deep excavations have become very common in South Korea. The average number of basement floors in buildings in South Korea, as measured by counting the number of buildings that appeared in the Korean Institute of Architects magazine [6], has increased from two to five since 1972. Recently, in Seoul, a retaining wall collapsed during deep excavation work, creating a 40-meter-long and 30-meter-deep crater. The pit not only swallowed five cars but also resulted in extensive damage to nearby buildings and underground utilities, causing power-supply outages and flooding. The inappropriate selection of a retaining wall system may not be the sole cause of this kind of serious failure, but it is common to find cases where the inappropriate selection of a retaining wall system has led to serious schedule delays and increases in costs due to unexpected changes in the construction method during construction.

Selecting a retaining wall system is a complex process, considering the various geotechnical and non-geotechnical factors involved. To help engineers choose a retaining wall system appropriate for a construction site, previous researchers proposed using machine learning techniques (a.k.a. data-mining techniques) based on excavation case histories [7,12–14,21–24]. In general, the machine learning technique requires thousands of cases to derive a reliable conclusion [4], but such a large number of excavation cases are very difficult to acquire in the construction domain. Consequently, the number of cases used in previous studies was usually smaller than 300. Another problem with previous studies is that the variables used in the machine learning techniques were chosen without a rigorous validation process for testing the correlations between the variables and the excavation methods. For these reasons, the application of (semi) automated machine learning techniques revealed several limitations. Therefore, this paper proposes to take a statistical approach, which provides researchers with more control over the selection and validation of variables and rule development based on a relatively small number of cases compared to the automated machine learning techniques. This paper first reviews previous studies with detailed examples and then explains the proposed statistical approach for building a decision tree for selecting retaining wall systems. Finally, using the same data set collected through this study, this paper compares the prediction accuracy rate (how accurately a prediction model can predict the outcome) of the decision tree developed in this study with that of another decision tree that is built using the Classification and Regression Trees (CART), a common machine learning algorithm.

2. Previous research using machine learning techniques

The machine learning technique aims to discover and generalize structural patterns in historical data. Although the theories of

* Corresponding author. Department of Architectural Engineering, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul 120-749, Republic of Korea. Tel.: +82 2 2123 5785; fax: +82 2 386 4778.
E-mail address: glee@yonsei.ac.kr (G. Lee).

0926-5805/$ – see front matter
doi:10.1016/j.autcon.2010.06.005

machine learning algorithms are complex, the applications are relatively simple. Statistics is the basis for machine learning technique theories. They have been widely applied in the construction domain. One area in which these techniques were deployed in construction is litigation- and management-related issues such as predicting the outcomes of construction litigation, contractor performance, and mediation [2,20,25]. The other area is the selection of construction methods appropriate for certain site conditions and others. Significant investigations have been carried out especially to develop prediction models for selecting retaining walls appropriate for construction site conditions [7,12,21–24] probably because the impact of problems caused by inappropriate excavation work on the entire construction schedule and costs is generally greater than that of other work in building construction. Table 1 summarizes the explanatory (or predictor) variables and target variables used in previous studies regarding the selection of retaining walls.

The machine learning techniques used in previous studies included the Expert System (ES), Neural Networks (NNs), Case-Based Reasoning (CBR), and Rule Induction (RI). The advantages and disadvantages of each machine learning technique are summarized in Table 2.

The previous studies had a couple of drawbacks. First, the explanatory variables were collected through literature review and interviews with geotechnical experts, but were used in machine learning without first undergoing a process that rigorously validated the correlation between the variables and the excavation methods.

In addition, many machine learning techniques, such as NNs, lack human-interpretability. This black-box characteristic makes it difficult for the modeler to validate and justify the final results. On the other hand, decision tree methods, including the RI method, have a distinct advantage over other machine learning techniques in that decision tree methods produce rules that are explicitly represented as a set of human-interpretable decision rules. The decision tree method exhaustively breaks down cases into a branched, tree-like form until the splitting of the data is statistically meaningful. Unnecessary branches should then be pruned using other test cases to avoid overfitting. This process generally requires a very large quantity of data (thousands of data) [9], which is rare, especially in construction.

To build a prediction model, a data set is split into three groups: a training set, a test set, and a validation set. A training set is used to build an initial model: i.e., a machine learning algorithm identifies a pattern in the training set through the training process, and the initial model is refined using a test set. Next, a validation set is used to evaluate the performance of the model with the prediction accuracy rate generally used as the performance index of a prediction model. If the value of the target variable given by the model matches the value of the target variable in the validation set, then the case is counted as "Correct." A few cases in the data set may consist of wrong decisions because even experts are not always right; this problem is relieved as the number of cases increases.

The appropriate number of cases for deriving a reliable conclusion varies according to machine learning technique. In the case of NNs, for example, the minimum number of data needed to effectively train a system is computed as follows: [4, p 159] in a fully connected neural network, the number of weights (i.e., the number of connections among the number n of input nodes, the number h of hidden nodes in a hidden layer, and the single output node) is calculated by $h\ (n+1)+1$. As a general rule of thumb, every weight requires at least 10 training instances, and when predicting categorical values, having this number of training instances for each category is advantageous. If we apply this calculation to an existing study conducted by Kim et al. [7], whose neural networks consisted of 15 hidden nodes, 10 inputs, and 7 categories as outputs, the study required at least 11,620 cases. However, the study was conducted based on 119 training cases. The largest data set used in the previous studies we reviewed is 254 (Table 3), which is still too small to get reliable results using automated machine learning techniques.

Table 2
The advantages and disadvantages of machine learning techniques applied in previous studies.

Techniques	Advantages	Disadvantages
Expert System (ES) [13,14]	Expert knowledge can be generalized into a set of rules	Knowledge elicitation is difficult and demanding; often sensitive to subjective experience
Neural Networks (NNs) [7,13]	Useful in detecting complex relationships between entities	No explanation or justification of decisions can be given, i.e., a "black box" characteristic
Case-Based Reasoning (CBR) [7,23]	Produces interpretable results; does not require a machine training process	Predictive indexes should be decided by a system developer (Nearest-neighbor retrieval)
Rule Induction (Decision Tree) [21,24]	Generates understandable rules	Requires a large number of training data; sensitive to relevance and type of variables

Table 1
Explanatory and target variables in previous studies.

Previous studies	Explanatory variables	Target variables
Yang [21] Yang [22] Yau and Yang [23] Yau et al. [24]	•Address (location) •Excavation depth (m) •Field area (m^2) •Working space •Pollution prevention •Neighboring settlement •Water table (m) •Soil type (nominal) •Soil strength (soft/firm/dense/loose) •Special soil condition (nominal)	•Slurry wall •Steel sheet pile •Steel rail pile •Steel pipe pile •H-section steel pile •Driven pile •Auger boring pile •Prepakt Mortar pile •Retaining column •Full casing pile •Row pile •Open excavation
Kim et al. [7] Park and Kim [12]	•Excavation area (m^2) •Excavation depth (m) •Shape of site (nominal) •Difference in the site level (m) •Number of sides that have buildings nearby •Distance to adjacent buildings (m) •Water table (m) •Thickness of soft soil layer (m) •Thickness of weathered rock layer (m) •Thickness of soft rock layer (m)	•Slurry wall •Soil cement wall •Cast-in-place concrete pile •Soldier piles and lagging •Soldier piles and lagging + jet grouting •Soldier piles and lagging + LW grouting •Soldier piles and lagging + soil cement wall