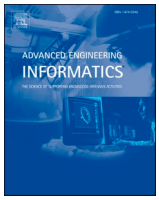
Advanced Engineering Informatics 49 (2021) 101354

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Converting detailed estimates to primary estimates with data augmentation

Yoshiyasu Takefuji

Musashino University, Faculty of Data Science, 3-3-3 Ariake, Koto-Ku, Tokyo 135-8181, Japan

ARTICLE INFO

Keywords:
Data augmentation
Primary estimate
Detailed estimate
Super-skilled engineers

ABSTRACT

In general, preliminary or primary cost estimates are used to select contractors from among bidders in Japan. The primary cost estimate must be accurate, otherwise the contractor selected from the bidding process will lose profit. A general contractor in the world does not have a super-skilled engineer who can achieve the accurate primary cost estimates. The conventional primary estimate has a high error range and low reliability. An automated system converting detailed estimates to primary estimates has been highly demanded in the world. This paper presents a prototype AI converter that can accurately and automatically convert detailed cost estimates into primary estimates. Converting detailed cost estimates to primary estimates lies in a regression problem. This paper proposes a feature-elimination based data augmentation method for regression problems. The empirical experiment shows that the proposed data augmentation method is quite effective with an Extra-Trees ensemble method. The proposed method was empirically examined by using Colorado Department of Transportation (CDOT) dataset for accurately predicting constructions costs with the Extra-Trees algorithm and random forest algorithm respectively. The CDOT dataset is one and only one of the largest datasets available in public for constructions costs quotation/estimation of roads, bridges and buildings.

1. Introduction

Although there are many construction cost estimate types, there are basically three types in general: design estimates, preliminary or primary estimates, and detailed estimates. In Japan, primary cost estimates are used to select contractors from among bidders due to the time-consuming nature of detailed estimates. Therefore, the primary cost estimate must be accurate, otherwise the contractor selected from the bidding process will lose profit.

Many general contractors in the world do not have super-skilled engineers who can achieve accurate primary cost estimates instead of detailed cost estimates. In other words, conventional primary estimates are unreliable because they have a large error range and rely solely on basic information about projects without super-skilled engineers [1,2].

In Japan, major general contractors have a few super-skilled engineers where they are very rare (one in 5000 employees), according to private communications with major construction contractors in Japan. Since preparing a detailed estimate is a time-consuming task, primary estimates have been used mainly in the construction industry and bidding system in Japan.

Many multinational construction companies and governments have shown a strong interest in efficient and accurate primary estimates. However, they do not have super-skilled engineers for accurate primary cost estimates. In order to introduce primary estimates into the bidding system, there is a strong need to convert detailed estimates into primary estimates.

This paper presents a prototype AI converter that can accurately and automatically convert detailed cost estimates into primary estimates. The primary estimates generated will tell us what are the important and dominant parameters that will determine the final cost.

The proposed method was empirically examined by using Colorado Department of Transportation (CDOT) dataset for accurately predicting constructions costs with the Extra-Trees algorithm and random forest algorithm respectively. The CDOT dataset is one and only one of the largest datasets available in public for constructions costs quotation/estimation.

The CDOT dataset is composed of 1458 instances (projects) and 8751 parameters for construction of roads, bridges and buildings. Although all materials used in roads, bridges and buildings in 1458 projects are included in the CDOT dataset, the detailed materials are not disclosed in the dataset.

When estimating the cost of a construction project, there are many factors to consider. Factors include: understanding the potential for missed scope (change orders), taking into account fluctuating construction material rates, historical costs and work histories of companies submitting proposals.

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.aei.2021.101354
Received 7 February 2021; Accepted 22 June 2021

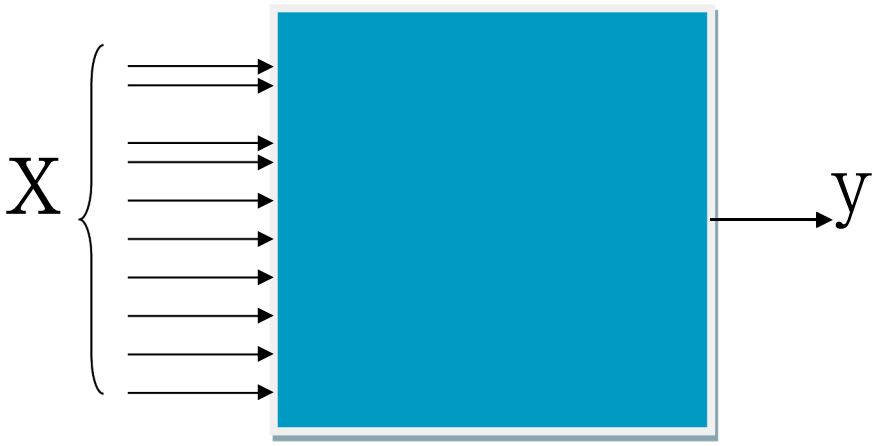


Fig. 1. Supervised machine learning: y = f(X).

Current estimating processes rely heavily on estimator knowledge and can be considered just as much art as science. The aim of CDOT project is to simplify and streamline the estimating process using machine learning.

Image data or text data augmentation techniques have been used for improving classification in machine learning. For example, Synthetic Minority Oversampling Technique (SMOTE) and/or Adaptive Synthetic sampling (ADASYN) have been popular methods for data augmentation classification with noise [3].

The goal of this paper is to build an accurate primary cost estimator from the detailed cost estimations and to show the effectiveness of the proposed data augmentation method by the empirical experiments using one of the largest datasets available in public. With the proposed method, a general contractor does not need a super-skilled human engineer for generating accurate primary estimates. In other words, the proposed method can select important parameters from massive parameters in the order of cost-contribution importances. Based on the computed accurate primary estimate, clients and constructors can discuss the critical cost issues and their decision-making can be fully supported.

The data augmentation methods play a key role in improving prediction accuracy in imbalanced data for solving classification problems. We call a dataset skewed or imbalanced if the majority of its data items represent items belonging to a certain class. Therefore, in the classification problem, we want to use the same number of instances of all classes in the training data for machine learning.

There is an important difference between classification and regression problems. Fundamentally, classification is to predict a label or a finite integer while regression is to predict a quantity or a real number. Therefore, the conventional data augmentation approaches (SMOTE or ADASYN) cannot be easily used for regression problems.

There is no accurate and efficient conversion from detailed estimates to primary estimates as far as we know. This paper shows the effectiveness of the proposed method for converting the largest detailed

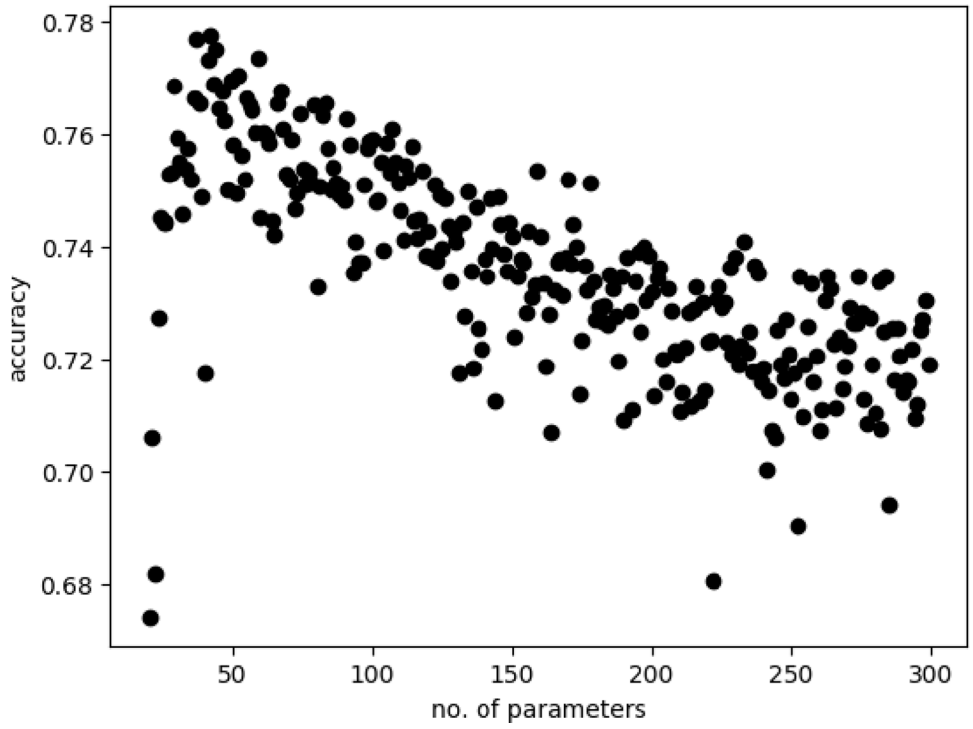


Fig. 2. Prediction accuracy vs the number of input parameters in Extra-Trees algorithm X-axis is the number of extracted parameters and Y-axis for prediction accuracy.

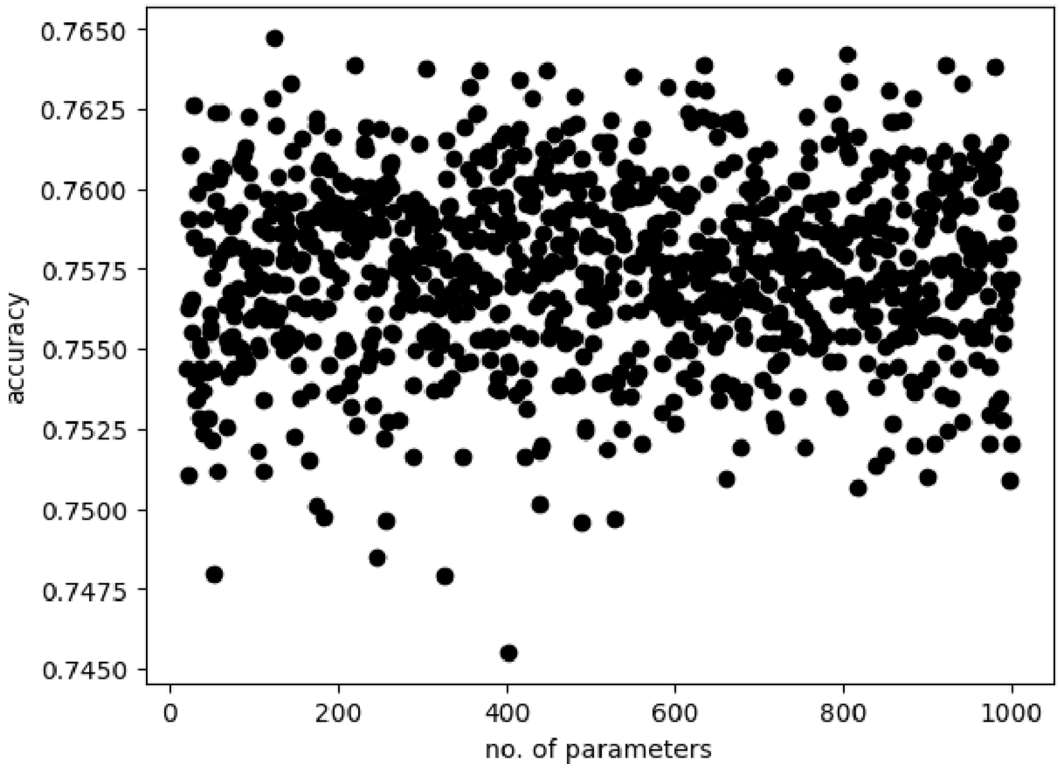


Fig. 3. Prediction accuracy vs the number of input parameters in random forest algorithm X-axis is the number of extracted parameters and Y-axis for prediction accuracy.

estimate using CDOT dataset into an accurate primary estimate.

2. Method

For solving regression problems, the proposed general-purpose data augmentation is based on feature-of-importance. The proposed data augmentation for regression is based on a simple feature-elimination method where parameters with less feature-importance can be eliminated while parameters with more feature-importance can be selected. Therefore, all input parameters are ordered along with feature-importance.

In Extra-Trees algorithm, features and splits are randomly selected so that the feature-of-importance indicates the positive influence against the target. In general, the target y is given by y = f(X) = f(x1, x2,…,xn) where x1, x2, and xn are input parameters and f() can be trained as the relationship function between input parameters and the target y as shown in Fig. 1.

The proposed data augmentation algorithm is described as follows:

1. Compute the feature of importance of all input parameters and reorder them from the viewpoint of importance values.
2. Extract input parameters based on the feature-of-importance from all input parameters. Use the extracted input parameters for training and test data for prediction accuracy.
3. Increment the number of input parameters by one.
4. Repeat 2 until the number of input parameters reaches the certain number.

3. Experiments and discussions

Random forest is a de facto ensemble algorithm implementing the wisdom of crowds. The random forest algorithm is composed of weak learners (trees) with a winner-take-all approach. The Extra-Trees algorithm and random forest algorithm were both examined by using CDOT dataset (1458 instances and 8751 parameters)[4] for data augmentation in regression. CDOT dataset was divided into train.csv (1093 instances) [5] and test.csv (365 instances)[6].

cdot_ext.py[7] and cdot_rf.py[8] programs in Python 3.7 were developed for examining the proposed general-purpose data augmentation in regression.

The result of cdot_ext.py in Extra-Trees algorithm shows that prediction accuracy = 0.71187 is obtained by using the maximum 8747 input parameters. Fig. 2 shows the relationship between the number of used input parameters and prediction accuracy in Extra-Trees algorithm. Fig. 2 shows that the best prediction accuracy = 0.77743 was obtained with data augmented only 42 input parameters. This means that successful primary cost estimate with 42 input parameters was achieved by the proposed data augmentation algorithm instead of using all 8747 input parameters. The successful primary cost estimates can play a key role in civil engineering business. Because the detailed cost quotations are time-consuming tasks. With the proposed data augmentation, accurate primary cost estimates can be achieved within a short time period.

In Japan, primary cost estimates play a key role in civil engineering business. In Japan, there is one and only one super-skilled engineer among 5000 engineers for accurate primary cost estimates. The accuracy of their primary cost estimates by the super-skilled engineer is the almost same as that of the detailed cost quotations in Japan. In China, there is no super-skilled primary cost estimate engineer. Therefore, the proposed system can play a key role in predicting primary cost estimates on behalf of the super-skilled engineers.

The result of cdot_rf.py in random forest algorithm shows that prediction accuracy = 0.7648 was obtained by using the maximum 8747 input parameters. Fig. 3 shows the relationship between the number of input parameters and prediction accuracy in random forest algorithm. The successful primary cost estimates can be achieved by less than 100 input parameters as shown in Fig. 3.

The main difference between random forests and Extra-Trees lies in

the fact that, instead of computing the locally optimal feature/split combination for the random forest for each feature under consideration, a random value is selected for the split for the Extra-Trees.[9] The proposed method is based on the globally optimal feature-split combination instead of locally optimal feature-split combination used in random forests. The experimental results can justify the proposed claim.

4. Conclusion

The proposed method for converting detailed estimates into primary estimates is based on the globally optimal feature-split combination which can improve the solution quality by Extra-Trees algorithm. In other words, the globally optimal feature is added to the existing Extra-Trees algorithm by eliminating unimportant features. Instead of the detailed estimate using the maximum 8747 input parameters with the prediction accuracy = 0.71187, the experimental results show that the proposed algorithm using data augmentation in Extra-Trees algorithm with less than 50 parameters can generate the accurate primary estimates with prediction accuracy of 0.77743 without losing prediction accuracy which is better than that of using the entire 8747 parameters. In other words, primary cost estimates using around 50 parameters were successfully generated by the proposed method without losing prediction accuracy. The proposed data augmentation not only shortens the cost quotations time but also provides the accurate prediction. From the detailed cost estimates, the proposed algorithm can generate the accurate primary cost estimates without a human expert involved. The proposed method will be useful for clients and constructors for cost estimates and decision-making when they would like to introduce primary estimates into the bidding system without a super-skilled engineer.

Funding

There is no fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jaewook Lee, Hyuncheul Yang, Jinkang Lim, Taehoon Hong, Jimin Kim, Kwangbok Jeong, BIM-based preliminary estimation method considering the life cycle cost for decision-making in the early design phase, Journal of Asian Architecture and Building Engineering 19 (4) (2020) 384–399, https://doi.org/10.1080/13467581.2020.1748635.
[2] Opeoluwa Akinradewo et al. (2020). Accuracy of road construction preliminary estimate: examining the influencing factors, Built Environment Project and Asset Management Vol. 10 No. 5, 2020pp. 657-671.
[3] M. Arslan, M. Guzel, M. Demirci and S. Ozdemir, “SMOTE and Gaussian Noise Based Sensor Data Augmentation,” 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 1-5.
[4] CDOT project estimator (2020). https://github.com/schustda/CDOT-Project-Estimator.
[5] CDOT train.csv file (2020). https://github.com/schustda/CDOT-Project-Estimator/blob/master/data/model/train.csv.
[6] CDOT test.csv file (2020). https://github.com/schustda/CDOT-Project-Estimator/blob/master/data/model/test.csv.
[7] CDOT extratrees python program cdot_ext.py (2020). https://raw.githubusercontent.com/ytakefuji/cdot/master/cdot_ext.py.
[8] CDOT randomforest python program cdot_rf.py (2020). https://raw.githubusercontent.com/ytakefuji/cdot/master/cdot_rf.py.
[9] RUser4512, (2018), Random-forest-vs-Extra-Trees, https://www.thekerneltrip.com/statistics/random-forest-vs-extra-tree/.