Automation in Construction 43 (2014) 23–29

Contents lists available at ScienceDirect

Automation in Construction

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

Predicting construction cost overruns using text mining, numerical data and ensemble classifiers

Trefor P. Williams *, Jie Gong [1]

Civil Engineering, Rutgers University, Department of Civil and Environmental Engineering, 96 Frelinghuysen Road, Piscataway, NJ 08854-8014, USA

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 31 January 2014
Accepted 22 February 2014
Available online 18 March 2014

Keywords:
Construction cost
Data mining
Text mining
Prediction

ABSTRACT

This paper discusses how text describing a construction project can be combined with numerical data to produce a prediction of the level of cost overrun using data mining classification algorithms. Modeling results found that a stacking model that combined the results from several classifiers produced the best results. The stacking ensemble model had an average accuracy of 43.72% for five model runs. The model performed best in predicting projects completed with large cost overruns and projects near the original low bid amount. It was found that a stacking model that used only numerical data produced predictions with lower precision and recall. A potential application of this research is as an aid in budgeting sufficient funds to complete a construction project. Additionally, during the planning stages of a project the research can be used to identify a project that requires increased scrutiny during construction to avoid cost overruns.

1. Introduction

Many factors affect construction cost overruns. Possibly, indicators provided in bidding text documents can identify construction projects that potentially will have large cost overruns. Text summaries of what is to be constructed for a project, and project line item text are available from project bidding data collected by state transportation agencies. Additionally, numeric data are available at the time of the bid opening including the projects' magnitude, and the number of bidders. It has now become possible to combine various text and data mining methods to project data to attempt to make predictions.

With the development of text mining algorithms that allow the extraction of information from the text, it may be possible to find indications of the projects' nature and likelihood to experience cost overruns. Text mining can be defined as the automatic discovery of previously unknown information from unstructured text data. Text mining involves extracting information of interest from text documents and then the use of data mining to discover new associations among the extracted information [1].

Text mining has been found to give excellent results for some predictive applications. For example, researchers have recently created software that detects when and where disease outbreaks will occur based on two decades of New York Times articles and other on-line sources of data. The system was successful in forecasting outbreaks of disease, violence and significant numbers of deaths between 70 and 90% of the time [2]. A construction project will typically have extensive documents produced before construction is initiated including project-scoping documents, written specifications, and plans. All of these documents can be transformed into text files useable by text mining software. In the construction industry, controlling costs is always a particular concern. Owner organizations require knowledge of a completed projects' cost to budget sufficient funds to complete a project. Additionally, knowledge of a project's likelihood to have cost overruns can identify projects that need to receive increased scrutiny during the construction process to control costs.

The goal of this research is to determine if text descriptions of a projects' characteristics can be used to develop a predictive model of a competitively bid construction projects' expected cost overrun. This paper will discuss how text can be processed, combined with numeric values and classified using data mining algorithms to produce a prediction of the level of cost overrun for a construction project.

1.1. Background

There have been several applications of text mining to construction management problems. Existing construction text mining research has focused on methods of classifying documents and extracting information from databases. A prototype system that automatically classifies construction documents according to project components using data mining techniques was proposed by Caldas et al. [3]. Soibelman and Kim [4] addressed the need for data mining in the construction industry, and the possibility to identify predictable patterns in construction data that were previously thought to be chaotic. In that study, a prototype knowledge discovery and data mining (KDD) system was developed to find the cause of activity delays from a U.S. Army Corps of Engineer's

* Corresponding author. Tel.: +1 848 445 2880.
E-mail addresses: tpw@rci.rutgers.edu (T.P. Williams), jiegong.cee@rutgers.edu (J. Gong).
[1] Tel.: +1 848 445 2881.

http://dx.doi.org/10.1016/j.autcon.2014.02.014
0926-5805/

database called the Resident Management System. Soibelman et al. [5] have addressed the need to develop additional frameworks that allow the development of data warehouses from complex construction unstructured data and to develop data modeling techniques to analyze common construction data types.

Various modeling techniques have been applied to the prediction of construction costs. They have usually focused on the use of numeric data to predict the project's outcome. Recent work has employed advanced data mining techniques to produce predictions. Son et al. [6] have developed a model using Principal Component Analysis and Support Vector Regression using 64 project definition variables to predict cost performance on building projects. Gritza and Labi [7] have applied econometric models to the analysis of highway project cost overruns. They found that, for a given project type and project duration, contracts of larger size or longer duration are generally more likely to incur cost overruns. Regression analysis and neural networks have also been applied to predicting construction costs [8–13]. Potentially, the addition of text data to various modeling techniques can enhance the predictions made by these models by covering more of the factors that can affect construction performance then can be derived from numeric data only.

2. Methods

2.1. Bidding data

Data for this analysis was collected from California Department of Transportation websites. Data from 1221 competitively bid highway projects were collected. The low bid, the completed project cost, and the numbers of bidders were collected. The percentage cost increase for each project was calculated as: Percentage Cost Increase = ((Completed Cost − Low Bid)/Low Bid) * 100.

Through experimentation, it was found that trimming large cost under run outliers from the data set produced better predictions. Therefore, 47 projects that were completed at a cost more than 25% lower than the original bid price were excluded from the analysis. The maximum cost overrun in the data set was 74% greater than the original low bid amount. Possibly, projects with very large cost under runs indicate a situation where the project scope was modified after the bid opening.

Each project was assigned to one of three cost overrun groups. Large cost overrun projects are categorized as having cost increases greater than 6%. Projects categorized as being completed near the original low bid had overruns between +6% and −3% under runs. Projects categorized as under run projects were all projects completed with an under run more than 3% less than the bid amount. The output of the model is a prediction of which of the three levels of overrun a project will have.

In addition to the numerical data, descriptive text was collected for each project. A short two to three sentence project description was obtained from a project summary included in the bid opening details. No location data was included (i.e. county or highway route). The text descriptions of the five largest project line items by dollar value were also collected. This data was added to include words in the analysis that describe the major work tasks on the project. These were the text describing standard unit price line items used by the California Department of Transportation. The text data for each project were saved as a continuous block of text. The project data collected varied widely in cost magnitude and type of construction. Some projects were maintenance projects while others were major rehabilitations or new construction. Table 1 shows an example of the input data.

2.2. Modeling software

The Rapid Miner software is a widely used data and text mining system [14]. The software incorporates powerful tools for data manipulation, data mining, and text mining. The Rapid Miner software allows experimentation with different types of data mining algorithms. In addition, the Rapid Miner software has the ability to manipulate and transform text into a format useable by data mining algorithms. All of the algorithms used in this research were originally developed for the Weka data mining software [15], and have been ported to the Rapid Miner system.

2.3. The modeling process

Various models were constructed that combined the text and numerical data to predict the level of cost overrun (or under run). Several different data mining algorithms were employed in the models with varying levels of success.

Fig. 1 illustrates the training process and testing processes. For each model type and model run 60% of the data were used for training (644 projects) and 40% was used for testing (430 projects). The first step in training is to split the data. The text data and numerical data are separated and processed separately. In the training process, the text data is submitted to text-mining algorithms to transform the text into a useable format and to provide data about the words and word pairs that are indicative of certain levels of cost overrun. The text for each project must be transformed into a numerical vector that is suitable for use with a data-mining algorithm. There are several steps that are necessary to transform the unstructured text for each project into a standardized numeric vector. These steps are tokenization, stopping, stemming, normalization and vector generation [16].

The output from the text processing is a very large sparse matrix with all of the words and word pairs as columns and projects as rows. Singular value decomposition (SVD) was used to reduce the text matrix to a smaller matrix of numerical data representing each project. This was necessary because the data mining models used can only accept numeric inputs and this transformation allows for faster computer processing.

The numerical variables are combined with the text data after the text data is processed and transformed into numeric variables. This combined data was submitted to a classification model. The models studied included Ripple Down Rules (Ridor), K-Star, Radial Basis Function (RBF) Neural Networks, and the Ensemble Stacking Method.

Table 1
Model input data.

Low bid	Number of bidders	Cost overrun level	Text
887859	3	3	INSTALL TRAFFIC SIGNAL SIGNAL AND LIGHTING ASPHALT CONCRETE (TYPE A) CLASS 2 AGGREGATE BASE 600 MM REINFORCED CONCRETE PIPE
42576814.6	4	3	BRIDGE CONSTRUCTION STRUCTURAL CONCRETE, BRIDGE TIME-RELATED OVERHEAD MOBILIZATION STRUCTURAL CONCRETE, BRIDGE FOOTING
1910418.36	4	3	RESURFACE ASPHALT CONCRETE RUBBERIZED ASPHALT CONCRETE (TYPE G) TRAFFIC CONTROL SYSTEM MOBILIZATION COLD PLANE ASPHALT CONCRETE PAVEMENT
1256988.37	4	3	UPGRADE PLANTING AND IRRIGATION PLANT (GROUP A) PLANT ESTABLISHMENT (LOCATION #1) (3 YEAR) PLANT ESTABLISHMENT (LOCATION #2) (1 YEAR) CLASS 2 AGGREGATE BASE
1996587.63	6	3	RESURFACE ROADWAY RUBBERIZED ASPHALT CONCRETE (TYPE G) REPLACE ASPHALT CONCRETE SURFACING COLD PLANE ASPHALT CONCRETE PAVEMENT TRAFFIC CONTROL SYSTEM