Journal of Applied Research on Industrial Engineering

Journal of Applied Research on Industrial Engineering

Vol. 2, No. 1 (2015) 15-33

Journal of
Applied Research on Industrial Engineering

Studying the suitability of different data mining methods for delay analysis in construction projects

Farzad Movahedi Sobhani [1] , Tahereh Madadi[2*]

[1] Department of Industerial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran (fmovahedi@iau.ac.ir)

[2] Master's Degree, Department of Industerial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran (p.madadi66@gmail.com)

ARTICLE INFO

Article history :
Received: 5 January 2015
Received in revised format: 12 April 2015
Accepted: 15 February 2015
Available online: 13 June 2015

Keywords :
Construction delay, data mining, evaluation, prediction, classification, factor selection

ABSTRACT

The main purpose of this paper is to investigate the suitability of diverse data mining techniques for construction delay analysis. Data of this research obtained from 120 Iranian construction projects. The analysis consists of developing and evaluating various data mining models for factor selection, delay classification, and delay prediction. The results of this research indicate that with respect to accuracy and correlation indexes, genetic algorithm with K-NN learning model is the most suitable model for factor selection. By conducting the genetic algorithm, eight significant variables causing construction delay are identified as: Changes in project manager, Difficulties in financing project by owner, Number of employees, Project duration, Unforeseen events, Project Location, Number of equipment, How to get the project. This research also revealed that in the case of delay classification and prediction, respectively, bagging decision tree and bagging neural network has the least amount of error in comparison with other techniques. In addition, to compare the diversity of data mining methods, the optimized parameter vectors of the selected models were also identified.

1. Introduction

In construction projects, delay is a universal phenomenon (Ahmed et al 2003; Assaf and Al-Hejji 2006; Frimpong et al 2003). It has negative effects on project parties (Ahmed et al 2003; El-Razek et al 2008); to the owner, it means loss of revenue; and to the contractor, it means higher overhead cost (Assaf and Al-

*Corresponding author name: Tahereh Madadi E-mail address: p.madadi66@gmail.com

Hejji 2006). Delays give rise to dissatisfaction of all involved parties, because it is often accompanied by high cost, low profitability, and poor quality (Aziz 2013; Odeh and Battaineh 2002).

Many articles and studies have been conducted on construction delay. For decades, a popular research stream has been committed in this area to identify and rank factors affecting delay (Assaf and Al-Hejji 2006; Doloi et al 2012), while, delay analysis and proposing ways to mitigate it has also been a research topic (Doloi et al., 2012; El-Razek et al 2008). Construction process is the subjected to many interrelated variables and unpredictable factors (Assaf and Al-Hejji 2006), making it difficult to discern the factors causing delay on a given project (Kim et al 2008). In addition, as many of these studies are area specific (Odeh and Battaineh, 2002), the applicability of such researches are in doubt.

In recent years, the volume of data in construction databases has grown enormously. The amount of data in construction databases contains large number of records, with many attributes of construction projects. These databases are mines of information and knowledge, which can be explored to discover useful knowledge of delay in construction industry. The diversity of data mining tools provides a great opportunity to select factors affecting delays as well as prediction of the amount of delay. One advantage of utilizing data mining techniques in the subject of construction delay is that the results are compatible with the different aspects of construction project management context. However, in spite of the rapid growth in the application of data mining in construction, there is still slow adoption of these techniques in the subject of delay. Particularly, this area faced with a big challenge. The diversity of data mining models provides a number of possible methods, but the appropriateness of these techniques for construction delay analysis is still unexplored. Thus the primary objective of this study is to determine the most appropriate model for cause of delay selection, and to find the most appropriate model for project delay prediction. Moreover, for each studied technique, the best parameters of the constructed models are estimated.

To achieve these objectives, we conducted this research based on the data gathered from 120 Iranian construction projects.

2. Literature Review

2.1. Causes of construction delay

Construction delay is defined as “the time overrun either beyond completion date specified in a contract, or beyond the date that the parties agreed upon for delivery of a project” (Assaf and Al-Hejji 2006; El-Razek et al 2008). Delay is also defined as “an act or event which extends required time to perform or complete work of contract manifests itself as additional days of work” (Zack 2003). Delays have negative effects on project parties (Ahmed et al 2003; El-Razek et al 2008) and give rise to dissatisfaction of all involved parties (Aziz 2013). Due to the relationship between time, cost, and quality (Munns & Bjeirmi 1996), it is often accompanied by high cost, low profitability, and poor quality as well (Aziz 2013).

Delay in construction projects has been certainly a significant topic for investigation for decades (Doloi et al 2012; Assaf and Al-Hejji 2006). One stream of research in this area is related to the subject of delay analysis (Doloi et al 2012). While, other stream is related to identifying attributes and factors affecting construction delays (Assaf and Al-Hejji 2006). A vast majority of these studies are area specific (Kim et al 2008, Lo et al 2006), making it difficult to be applied by many countries. Below, certain literatures focusing on factors and variables causing delay are reviewed.

Aziz (2013) investigated the factors perceived to affect delay in Egyptian construction projects. A structured questionnaire was developed and distributed among practitioner and experts, through which

ninety-nine factors causing delay in construction projects are determined. Then, based on the quantified relative importance indices, he classified explored factors into nine primary categories as follows: (1) consultant related delay factors, (2) contractor related delay factors, (3) design related delay factors, (4) equipment related delay factors, (5) external related delay factors, (6) labor related delay factors, (7) material related delay factors, (8) owner related delay factors, (9) project related delay factors. Doloi et al. (2012) investigated key factors impacting delay in Indian construction industry. They adopted a questionnaire survey to find impact of various attributes on delay. A factor analysis and regression modeling was used to examine the significance of delay factors. From the factor analysis, most critical factors of construction delay were identified as: (1) lack of commitment, (2) inefficient site management, (3) poor site coordination, (4) improper planning, (5) lack of clarity in project scope, (6) lack of communication, (7) substandard contract. The regression model developed by Doloi (2009) indicates that slow decision from owner, poor labor productivity, architecture's reluctance to change, and rework due to mistakes are the reasons that affect the delay. Assaf and Al-Hejji (2006) investigated causes of construction delay and their importance in Saudia Arabia. They conducted a survey including 23 contractors, 19 consultancies, and 15 owners, on time performance of different types of construction projects. The outcome of the study indicates the "change order" as the most common cause of delay identified by all three parties. El-Razek et al. (2008) repeated this study in Egypt to identify main cause of delay in construction project from the point of view of contractors, consultancies, and owners. The overall results indicated that the most important causes are: financing by contractor during construction, delays in contractor's payment by owner, design change by owner, partial payments, and non-utilization of professional management. They also found that the contractor and owner have opposite views, mostly blaming one another for delay. Fallahnejad (2013), investigated delay causes in 24 Iranian transmission gas pipeline projects. In order to explore the common delay causes, the author reviewed related literatures, assessed previous projects, documents, and conducted some initial interviews with oil & gas experts. The outcome of this stage revealed 43 well-assessed delay factors categorized in 9 groups. Then, the author conducted a questionnaire survey and statistical tests and analyses to determine the importance of each item. The questionnaire survey shows that 10 most important delay causes are related to imported materials, unrealistic project duration, client-related materials, land expropriation, change orders, contractor selection methods, payment to contractor, obtaining permits, suppliers, and contractor cash flow. Frimpong et al. (2003) investigated causes of delay and cost overrun in construction project in Ghana. They conducted questionnaire survey to identify and evaluate the relative importance of the significant factors contributing to delay and cost overruns in Ghana groundwater construction projects. The results of the study revealed the main causes of delay and cost overruns in construction of groundwater projects including monthly payment difficulties from agencies, poor contractor management, material procurement, poor technical performances, and escalation of material prices. Odeh and Battaineh (2002) presented the findings of a survey aimed at identifying the most important causes of delay in construction projects from the viewpoint of contractors and consultants. Results of the survey indicated that consultants and contractors agreed that owner interference, inadequate contractor experience, financing and payments, labor productivity, slow decision making, improper planning, and subcontractors are among the top ten most important factors. Sambasivan and Soon (2007) conducted a questionnaire survey to solicit the causes and effects of delay from clients, consultants, and contractors in Malaysia. This study identified 10 most important causes of delay from a list of 28 different causes and 6 different effects of delay. Ten most important causes were: (1) contractor's improper planning, (2) contractor's poor site management, (3) inadequate contractor experience, (4) inadequate client's finance and payments

for completed work, (5) problems with subcontractors, (6) shortage in material, (7) labor supply, (8) equipment availability and failure, (9) lack of communication between parties, and (10) mistakes during the construction stage. Six main effects of delay were: (1) time overrun, (2) cost overrun, (3) disputes, (4) arbitration, (5) litigation, and (6) total abandonment.

The mentioned literatures are a few examples of numerous researches published in this regards. Although we did not present all the reviewed articles, the selected literatures are the most cited ones, containing famous previous works on causes of delays. From the literature review, the following points are noticeable:

- Almost all researchers adopted questionnaire survey to identify causes of delay and to identify their importance according to project's client, contractors, and consultants view.
- According to the comparative analysis of delay causes in various countries illustrated in Table 1, the causes of construction delays are broadly divided into two categories- one category including attributes and factors is common among various countries, and the second category is related to location specific factors.
- The applicability of these factors in other countries is in doubt. Moreover, these factors get their roots in construction industry context including project management practice, project PEST environment (political, economic, social and technological factors) and so on. Usually, these contextual factors are intertwined and dynamic, making it difficult to identify or predict cause of construction delays.
- The same story happens to the degree of importance of these factors; therefore ranks of these factors in given country vary from the others.

In spite of identifying various factors contributing to construction delays, fewer works have been published on prediction delay in a construction project.

2.2. Data mining

For the time being, by growing the volume of data collected in construction project database and data warehouse, data mining would be an alternative solution to identify causes of delays. Data mining is the non-trivial process of automatically discovering valid, novel, and useful information in large data repositories (Chi et al 2012; Sharma and Osei-Bryson 2009; Frawley et al 1992). The existing literature reveals a broad range of application of various data mining techniques in project management fields (Cheng et al 2010; Chua et al 1997). Examples include project performance prediction (Chi et al 2012), time management (Shadrokh and Aghdashi 2012), cost performance prediction (Son et al 2012), time performance prediction (Chan et al 2001), profit prediction (Han et al 2007), project success prediction model (ko and Cheng 2007), and factor selection for delay analysis (Kim et al 2008). However, fewer researchers have worked on utilizing data mining techniques to identify cause of construction delays. Table 2 illustrates different works focused on application of data mining in project management.

In this research, a series of data mining methods are used. glossaries of these methods are given in the following:

2.2.1. Decision tree

Decision tree is one of the most popular classification techniques. It was invented independently by Quinlan (known as ID3) and Breiman (known CART) at around the same time (Han and Kamber 2006:

292). Decision tree recursively partitions the records in the training data set into subsets of records with similar values for the target attributes (Larose 2005: 109). Both ID3 and CART adopt a top down recursive divide-and-conquer approach to generate decision tree. The input consists of three parameters: (1) data partition, D, which is a set of training tuples and their associated class labels; (2) attribute list, the set of candidate attributes; (3) attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes.

2.2.2. K-Nearest Neighbor

The k-nearest-neighbor method was first described in the early 1950s. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. "Closeness" is defined in terms of a distance metric, such as Euclidean distance (Han and Kamber, 2006: 348).

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (1)$$

2.2.3. Neural Network

Neural networks are used for classification or estimation. It compromises of a series of independent processors or nodes. These nodes are connected to other nodes and are organized into a series of layers. Multilayer feed-forward networks are one of the most important and most popular classes of artificial neural networks in real-world applications. Typically, the network consists of a set of inputs that constitute the input layer of the network, one or more hidden layers of computational nodes, and finally an output layer of computational nodes. The processing is in a forward direction on a layer-by-layer basis. This type of artificial neural networks is commonly referred to as multilayer perceptron (Kantardzic 2011).

2.2.4. Bayesian classification

Bayesian classifier, known as the naïve Bayesian classifier, is based on Bayes' theorem. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis, such that the data tuple X belongs to a specified class C. For classification problems, the objective is to determine P(H|X), the probability of that the hypothesis H holds, given the "evidence" or observed data tuple X (Kantardzic 2011: 311). P(H|X) is the posterior probability of H conditioned on X . The posterior probability is estimated as follow:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

3. Research Method

Various data mining methodologies have been proposed in the literature to provide guidance towards the process of implementing data mining projects (Sharma and Osei-Bryson 2009). Based on our review

of various data mining methodologies proposed in the literature (Anand and Buchner 1998; Berry and Linoff 1997; Cabena 1998; Cios and Kurgan 2005; CRISP-DM 2003; Fayyad et al 1996), we selected CEISP-DM methodology due to three reasons. First, this methodology is popularly used in real world organizations (Sharma and Osei-Bryson 2009). Second, it is more detailed than any other DM methodologies. Third, the steps of the methodology are independent from types of data (Han and Kamber 2006). CRISP-DM consists of six phases, shown in figure 1. The sequence is not rigid, moving back and forth between different phases.

According to different steps of CRISP-DM, we develop a framework for our research. The remains focus on steps defined in figure 2.

3.1. Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. Business understanding consists of four tasks: determination of business objective, situation assessment, determination of data mining goals, production of project plan.

Table 1. Delay causes in various countries

	Hong Kong (Chan & Kumaraswamy)	Jordan (Al-Momani 2000)	Malaysia (Sambasivan & Soon 2007)	Egypt (Aziz 2013)	UAE (Faridi& El Sayegh 2006)	Turkey (Arditi et al. 1985)	U.S. (Baldwin 1971)	Nigeria (Mansfield et al. 1994)	Zambia (Kaliba et al 2009)	Iran (Fallahnejad 2013)	Ghana (Frimpong et la 2003)	Florida (ahmed et al 2003)	Indonesia (Kaming et al 1997)	Indian (Doloi et al 2012)
Inadequate resources	*		*		*	*	*	*	*		*		*	*
Unforeseen ground condition	*						*	*		*				
Exceptionally low bids	*			*										
Inexperienced contractor	*	*	*		*					*				*
Works in conflicts with existing utilities	*													
Poor site management	*	*	*	*	*			*	*		*			*
Unrealistic contract duration	*									*		*		*
Environment restriction	*													
Slow coordination	*			*						*		*		
Change scope	*			*		*								*
Owner interference		*								*				
Improper payment		*	*	*	*	*		*	*	*	*	*		
Labor productivity		*			*								*	*
Slow decision making		*		*	*							*		*
Construction methods		*												
Improper planning subcontractor		*	*				*	*		*		*		
Equipment availability and failure			*	*						*				
communication			*											
Mistakes in construction phase			*						*	*				
Preparation and approval of drawings				*	*				*	*				*
organizational deficiencies						*						*		
considerable additional work						*				*				
inaccurate cost estimates								*			*			
Financing by contractor during construction				*										*
Design changes													*	
Delay in material delivery by vendors														*

Table 2. Data Mining in Project management

ID	Author name/Year	Purpose	Data	Technique	Country
1	Hongqin Fan et al (2008)	estimating the residual value of heavy construction equipment	Last BidTM, an online construction equipment database covering up-to-date auction results across the U.S. and international markets	Auto Regressive Tree Algorithm	U.S & Canada
2	Chang & Leu (2006)	a data mining model and procedure to relate influence variables to project profitability	548 projects of an engineering consulting company	The model similar to a data mining process	-
3	Cheung et al (2006)	Predicting project performance	information on the contractors performance are investigated based on the project data collected in Hong Kong	neural networks	Hong Kong
4	Cheng et al (2009)	to achieve strategic control over project cash flows	52 construction projects in Taipei	K-means clustering, genetic algorithm (GA), fuzzy logic (FL), and neural network (NN)	Taipei
5	Chua & Log (1997)	identifies key project management attributes associated with achieving successful budget performance	75 construction projects	Neural Network (NN)	-
6	Shahram Shadrokh(2012)	support project management by presenting various informative insights, which would enable the better understanding of the past dynamics and provide grounds for better planning of the future research project programs	The project from a construction company (kayson)	Clustering & Association rules	Iran
7	Iranmannesh & Mokhtari (2008)	predicting the total project duration in term of Time Estimate At Completion-EAC	sample project from Kulish – Hartmann data set (j303_10) from http://129.187.106.231/psplib/main.html.	Association rules	-
8	Cheng et al (2010)	Project success prediction	real data collected by Russell from 16 company members of the Construction Industry Institute (CII) (46 construction Projects)	support vector machine (SVM) & a fast messy genetic algorithm	-
9	Ma et al (2008)	Developing A decision support system for utilize the exchanged documents to support decision making of the management staffs in construction project	the main stadium for the Beijing 2008 Olympic Games	Decision Tree & Clustering	China
10	Son et al (2012)	Propose and validate a hybrid predictive model for cost performance of commercial building projects	84 sets of data from an equal number of commercial building projects	integrates a support vector regression (SVR) model with principal component analysis (PCA).	South Korea
11	Shahrabi & Taghavi (2012)	analyze the macroeconomic performance of different Mining and Industrial Projects of Iran	60 projects	Clustering & supervised learning techniques	Iran
12	Kim et al (2008)	presents a methodology for factor selection; identifying which factors in an on-going construction project contribute most to the experienced delays	A specific project in the Resident Management System (RMS). RMS is a large database used by the US Army Corps of Engineers to track construction contracts and progress information.	neural networks/Bayesian networks/Decision tree	US

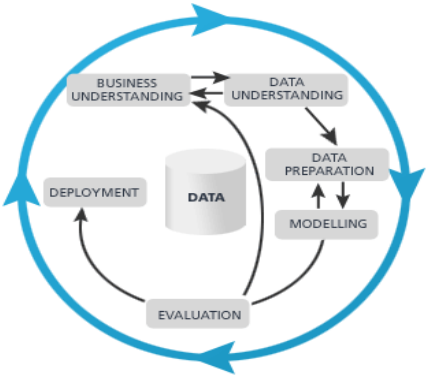

Fig. 1. Crisp-DM process model (Han and Kamber 2006)

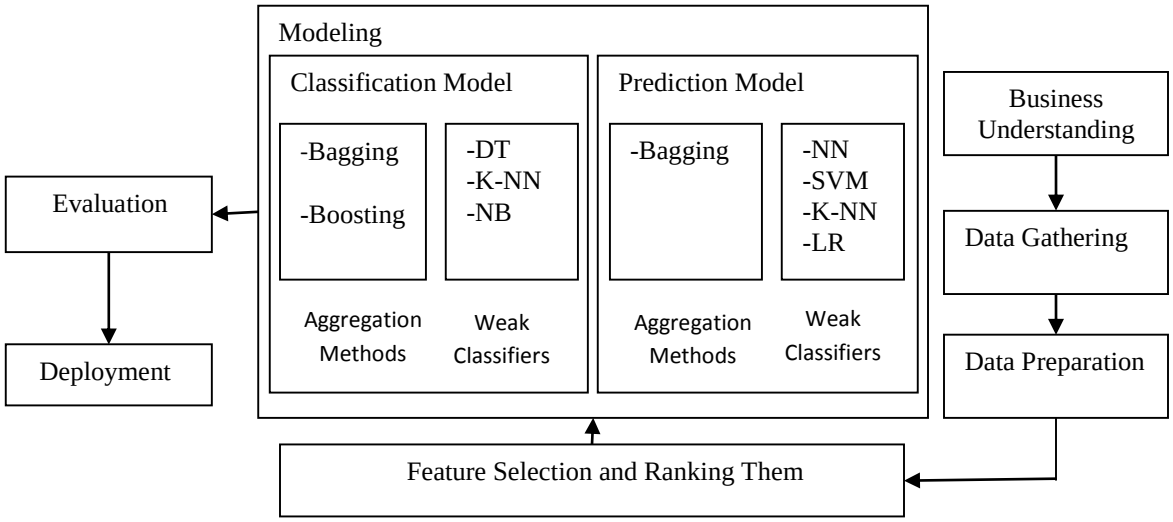

Fig 2. Research framework

3.2. Data understanding

Copies of database from sixteen construction companies were used in this research. The databases contained data on 120 diverse construction projects. Some selected projects were pure construction project, while others were relating to construction phase of larger projects in particular industry. Table 3 shows various types of selected projects.

Table 3. Types of selected projects

code	Project type
1	Water supply
2	Petrochemical
3	Water treatment
4	Wastewater
5	Manufacturing
6	Railway
7	Road

Table 3.	Types of selected projects
8	House building
9	Dam
10	Marine industry
11	Mineral industry
12	Oil& Gas

Based on our literature review as well as the availability of the related data in the databases, 17 variables were defined in this research. Table 4 shows variable names, measurement scale, and description.

3.3. Data preparation

The collected project data include noisy, missing, and inconsistent data. In the third phase, the quality of data needs to be improved and transformed into appropriate formats. In the case of missing data, it was substituted by relating mean.

Table 4. Variable names and description

Column no.	Variable name	Measurement scale	Description	Reference
1	Name	Nominal		
2	Contract type	Nominal	E:engineering, P: procurement, C: construction, Any combination of E,P,C	
3	Value	Ratio		
4	Start date	Nominal		
5	Duration	Ratio		Marzouk et al (2014)
6	Delay amount	Ratio	The difference between planned schedule and actual schedule	Kim et al (2008)
7	Financial situation	Ratio	Profit, loss, Breakeven	
8	Location	Nominal		Ahmed et al (2003) Marzouk et al (2014) Odabaşı (2009)
9	Number of employees	Ratio	Total man power per month	Baldwin et a (1971) Arditi et al (1985) Faridi et al (2006)
10	Number of equipment	Ratio	Total number of equipment per month	Baldwin et al (1971) Arditi et al (1985) Faridi et al (2006)
11	Contractor name	Nominal		
12	Customer type	Nominal	Government, private sector	
13	Project type	Nominal	Refer to Table 3	
14	Unforeseen events[1]	Nominal	1:Additional works, 2: Strike, 3: Weather effect (hot, cold), 4: Error of estimation in initial rates, 5: Fluctuation in exchange rate, 6: flood, 7: Hitting the water pipes, 8: Change of plans, 9: Fluctuation in cost, 10: Landslides, 11: Dust, 12: Failure to employer responsibilities	Muya et al (2009) Marzouk et al (2014) Al-Momani (2000)
15	Difficulties in financing	Ratio	The difference between the number of submitted invoices by contractors and the	Odabasi (2009) Muya et al (2009)

Table 4. Variable names and description

Column no.	Variable name	Measurement scale	Description	Reference
	project by owner		number of bills paid by owner in project life cycle	Al-Khall and Al-Ghafly (1999)
16	How to get the project	Nominal	(Tender/ Partnerships with other contractors/ Agreement with the employer)	Assaf et al (1995) Al-Khall and Al-Ghafly (1999) Marzouk et al (2014)
17	Changes in project manager	Ratio	Frequency of project management change in project life cycle	Mansfield et al (1994) Okpala and Aniekwu (1988)

3.4. Factor (feature) selection

As mentioned in literature review, factors causing construction delays are many in number. Moreover, much of them are correlated as well as redundant. In order to achieve maximum performance and higher accuracy from data mining, a process of feature reduction is often necessary. This process partitions the feature subsets into core features (strongly relevant), weakly relevant features, and irrelevant features. Kohavi and Sommerfield (1995) described the problem of feature subset as heuristic search. They introduced compound operators that dynamically change the typology of the search space to better utilize the information available from the evaluation of feature subset. The compound operator is based on genetic algorithm, forward and backward statistical method, machine learning, decision tree, and Bayesian classification. Yang and Honavar (1997) stated genetic algorithm could offer an alternative approach to find near-optimal solution to feature selection. In this research, we use wrapper method to find relevant features. The learning model in wrapper approach is based on four algorithms, including decision tree, neural network, Bayesian classification, and K-nearest neighbor.

3.4.1. Results

In general, data mining analysis involves two top-down steps: (1) structure identification, and (2) parameter identification. Structure identification focuses on determining the most suitable models; this step usually is conducted based on prior knowledge about target system. While the target of parameter identification is to apply optimization techniques in order to determine parameter vector. Table 5 illustrates the parameters of four data mining techniques that are conducted for factor selection.

Table 5. Employed models for factor selection

Model	Parameters	Abbreviation
Genetic Algorithm	Population Size:5 Probability Initialize (P_i): 0.5 Probability Mutation (P_m): 0.66 Probability Crossover (P_c):0.5 Selection scheme: Tournament	GA
Forward	Maximal number of attribiute:15	-
Backward	Maximal number of eliminations:10	-

Table 5. Employed models for factor selection

Model	Parameters	Abbreviation
Decision Tree C4.5	Criterion: Gain-Ratio Maximal Depth:30 Confidence: 0.25 Minimal Gain:0.1	DT
K-NN	K=1 Measure Types: mixed Euclidean Distance	-
Naive Bayes	-	NB
Artificial Neural Network	Learning Rate: 0.3 Training Cycle: 500 Momentum: 0.5 Hidden Layers:1	NN

For appraising the appropriateness of these models, the accuracy and correlation indexes for discrete and continuous target variables are calculated respectively. As it can be seen from Table 6, genetic algorithm with K-NN learning model is the most suitable model for factor selection.

Table 6. Evaluation of employed models for factor selection

ID	Model	Learning Model	Accuracy	Correlation	Number of selected variables
1	Forward	DT	88.60%	-	6
2	GA	KNN	100%	1.000	8
3	Forward	KNN	74.56%	0.905	3
4	GA	DT	88.60%	-	9
5	GA	NB	92.11%	-	10
6	Forward	NB	92.11%	-	7
7	Backward	NB	92.11%	-	12
8	Backward	KNN	79.82%	0.905	10
9	Backward	DT	84.21%	-	12
10	Backward	NN	94.74%	0.996	14
11	GA	NN	95.61%	0.998	11
12	Forward	NN	93.86%	0.991	7

By conducting genetic algorithm, eight variables possible for causing construction delays are selected. They are illustrated in Table 7. The amount of 100 percent for accuracy index and the correlation of 1 indicate that the eight selected variables are perfectly suitable for classifying and predicting delays in construction project.

Table 7. Selected variables

ID	Variable name
1	Duration
2	Location

Table 7. Selected variables

ID	Variable name
3	Number of employees
4	Number of equipment
5	Difficulties in financing project by owner
6	How to get the project
7	Unforeseen events
8	Changes in project manager

In addition to identifying factors affecting construction delays, the relevance of these factors are also of significance. The relevance shows that in what extent each factor contributes to the amount of given construction project. To achieve this purpose, we used the information gain ratio. This measure is based on pioneering work by Claude Shannon on information theory, who studied the value or "Information content" of messages (Han and Kamber 2006: 297).

The result of conducting information gain ratio is summarized in Table 8. As it can be seen "change in project manager" and Difficulties in financiering project by owner" are two factors with high impact on construction delays.

Table 8. Variables ranks based on "Information gain ratio"

ID	Variable	Weight
1	Changes in project manager	1.0
2	Difficulties in financing project by owner	0.936
3	Number of employees	0.677
4	Duration	0.444
5	Unforeseen events	0.363
6	Location	0.203
7	Number of equipment	0.150
8	How to get the project	0.0

3.5. Construct data mining model

For the purpose of this paper a two stage data mining model was constructed.

3.5.1. Stage 1: Delay classification

Project managers sometimes interest in predicting the amount of construction delay, while it is considered as a categorical variable. In this case the proper data mining approach is classification. Classification is a process of learning a function that maps a data item into one of several predefined classes. The aim of stage 2 is to create a classification model, which would assign a discrete label value to construction delays. Different approaches are defined for classification, in this research we applied four classification methods including K-Nearest Neighbor, decision tree, Bayesian classification, bagging and

boosting techniques. For the purpose of classification, the continuous value of delay discretized into four intervals, as illustrated in Table 9.

Table 9. Discretization of the target variable (Delay)

ID	Interval Delay(month)	Class Name	Definition
1	[-∞,0]	0	Acceleration
2	[1,10]	1	Delay
3	[11,20]	2	much delay
4	[30,+∞]	3	Too much delay

3.5.2. Results

Table 10 illustrates the parameters of different data mining techniques that are conducted for classification. All these models are conducted according to the eight variables which are selected through genetic algorithm.

For appraising the appropriateness of obtained models, they are compared with respect to three indexes such as accuracy, root mean square error, and absolute error. These indexes are calculated by using training data. Training data is obtained by splitting the data set into 80% training and 20% test samples. The result of evaluation is summarized in table 11.

Table 10. Parameters used in the classification techniques

Model	Parameters	Abbreviation
Decision Tree C4.5	Criterion: Gain-Ratio Maximal Depth:30 Confidence: 0.25 Minimal Gain:0.1	DT
Naive Bayes	-	NB
K-NN	K=10 Measure Types: mixed Euclidean Distance	-
Boosting- Decision Tree C4.5	Iterations:10	Ad-DT
Boosting- Naive Bayes	Iterations:10	Ad -NB
Boosting- K-NN	Iterations:10	Ad- K-NN
Bagging- Decision Tree C4.5	Iterations:10 Sample Ratio:0.9	B-DT
Bagging- Naive Bayes	Iterations:10 Sample Ratio:0.9	B-NB
Bagging- K-NN	Iterations:10 Sample Ratio:0.9	B-K-NN

Table11. Classification models evaluation results

ID	Model	Accuracy	Root mean square error	Absolute error
1	DT	78.26%	0.472	0.340
2	K-NN	52.17%	0.649	0.630
3	NB	56.52%	0.631	0.454
4	Ad- K-NN	52.17%	0.730	0.729
5	Ad- NB	56.52%	0.676	0.662
6	Ad- DT	78.26%	0.413	0.235
7	B- NB	60.78%	0.628	0.461
8	B- DT	82.61%	0.431	0.341
9	B- K-NN	47.83%	0.649	0.635

Table 11 indicates that among the obtained models, bagging decision tree has the best value for accuracy. While, with respect to root mean square error, boosting decision tree is the best. However, the relative deference between the accuracy and error of these two models indicates that bagging decision tree is more perfect than others. On the contrast, with respect to the calculated measures, bagging K-NN and boosting K-NN are the worst for delays classification.

3.5.3. Stage 2: Delay Prediction

The aim of stage 2 is to build data mining models for predicting continuous amount of delay. These models help project managers to make an estimation of project duration, while factors affecting construction delays are active.

To achieve this purpose, four data mining models including neural network, support vector machine, K nearest neighbor, and regression are used. In addition to overcome a particular weakness in learning method, backing strategy is also employed.

3.5.4. Results

Table 12 illustrates the parameters of different data mining techniques that are conducted for prediction. All these models are conducted according to the eight variables which are selected through genetic algorithm.

For appraising the appropriateness of obtained models, they are compared with respect to three indexes i.e. correlation, root mean square error, and absolute error. The indexes are calculated using training data. The results of evaluation are summarized in table 13.

Table 13 indicates that among the obtained models, regression and bagging regression has the best value for correlation. While, with respect to root mean square error, bagging neural network is the best.

We also calculate "mean absolute error" in order to measure the power of each prediction models.

That is,

$$Mean\ absolute\ error = \frac{Absolute\ error}{average\ delay} * 100 \qquad (3)$$

The results are summarized in table 14 (the average delay is 260.58 days).

Table12. Parameters used in the prediction techniques

Model	Parameters	Abbreviation
Artificial Neural Network	Learning Rate: 0.3 Training Cycle: 500 Momentum: 0.5 Hidden Layers:1	NN
Support Vector Machine	Kernel Type: anova C:0.25	SVM
K-NN	K=23 Measure Types: mixed Euclidean Distance	
Linear Regression	Feature Selection: T-Test Alpha:0.05	LR
Bagging- Artificial Neural Network	Iterations:10 Sample Ratio:0.9	B -NN
Bagging-K-NN	Iterations:10 Sample Ratio:0.9	B- K-NN
Bagging-Support Vector Machine	Iterations:10 Sample Ratio:0.9	B-SVM
Bagging-Linear Regression	Iterations:10 Sample Ratio:0.9	B-LR

Table13. Prediction models evaluation results

ID	Model	Correlation	Root mean square error	Absolute error
1	NN	0.8	4.832	3.598
2	SVM	0.571	8.715	7.803
3	K-NN	0.366	6.857	5.335
4	LR	0.934	6.178	4.662
5	B-NN	0.839	4.008	2.903
6	B- SVM	0.599	7.531	6.371
7	B- K-NN	0.319	6.907	5.280
8	B- LR	0.931	5.797	4.042

Table14. The results of the mean absolute error in prediction models

ID	Model	Mean absolute error
1	NN	41.42%
2	SVM	89.83%

Table14. The results of the mean absolute error in prediction models

ID	Model	Mean absolute error
3	K-NN	61.42%
4	LR	53.67%
5	B-NN	33.42%
6	B- SVM	73.34%
7	B- K-NN	60.78%
8	B- LR	46.53%

With respect to mean absolute error, bagging neural network has the least amount of error in comparison with other techniques. On the contrast, support vector machine, Bagging support vector machine, and K-NN are the worst techniques for delays prediction.

4. Conclusion

Both project managers and researchers are interested in exploring useful knowledge of delay from construction databases. They are interested in identifying factors affecting delays as well as predicting the amount of delay in construction projects. However, the appropriateness of the diversity of data mining techniques in the subject of construction delay is still unexplored. This research presented the results of comparing different data mining techniques, and proposed the best data mining methods in this regards.

Factors causing construction delay are many in number; therefore, at the first step of data mining process, it is necessary to employ the feature reduction to identify the most important uncorrelated factors. This research reveals that among different employed factor selection techniques, with respect to accuracy and correlation index, the genetic algorithm with K-NN learning model is the most suitable one. Based on the identified model's parameters, 8 factors which are of significance in classifying and predicting construction delays are explored.

By classification, a discrete label value is assigned to construction delays. This research reveals that among different employed approaches of classification, bagging decision tree has the best value for accuracy. While, with respect to root mean square error, boosting decision tree is the best. But if we consider the relative difference between the accuracy and error, bagging decision tree is the most perfect model for classification.

For delay prediction, the result of this research reveals that with respect to mean absolute error, bagging neural network has the least amount of error in comparison with other techniques.

By this research, the authors have attempted to study different data mining methods in construction delays and estimate the selected model parameters based on Iranian construction industry. In future researches, we propose to study the applicability of the obtained models in other countries.

5. References

Abd El-Razek, M. E., Bassioni, H. A., & Mobarak, A. M.(2008). "Causes of delay in building construction projects in Egypt", *Journal of Construction Engineering and Management*, Vol. 134, No. 11, pp. 831-841.

Ahmed, S. M., Azhar, S., Castillo, M., & Kappagantula, P. (2002). "Construction delays in Florida: An empirical study", *Final Report Submitted to State Florida, Department of Community Affairs*.

Al-Momani, A. H. (2000). "Construction delay: a quantitative analysis", *International journal of project management*, Vol. 18, No. 1, pp. 51-59.

Anand, S. S., & Büchner, A. G. (1998) *Decision support using data mining*, Financial Times Management.

Arditi, D., Akan, G. T., & Gurdamar, S. (1985). "Reasons for delays in public projects in Turkey",*Construction Management and Economics*, Vol. 3, No. 2, pp. 171-181.

Assaf, S. A., & Al-Hejji, S. (2006). "Causes of delay in large construction projects". *International journal of project management*, Vol. 24, No. 4, pp. 349-357.

Aziz, R. F. (2013). "Ranking of delay factors in construction projects after Egyptian revolution", *Alexandria Engineering Journal*, Vol. 52, No. 3, pp. 387-406.

Baldwin, J. R., Manthei, J. M., Rothbart, H., & Harris, R. B. (1971). "Causes of delay in the construction industry", *Journal of the Construction Engineering*,Vol. 97, No. 2, pp. 177-187.

Berry, M. J., & Linoff, G. (1997) *Data mining techniques: for marketing, sales, and customer support*, John Wiley & Sons, Inc .

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998) *Discovering data mining: from concept to implementation*, Prentice-Hall, Inc

Chan, A. P., Ho, D. C., & Tam, C. M. (2001) "Design and build project success factors: multivariate analysis", *Journal of construction engineering and management*, Vol. 127, No. 2, pp. 93-100.

Chan, D. W., & Kumaraswamy, M. M. (1997). "A comparative study of causes of time overruns in Hong Kong construction projects", *International Journal of project management*, Vol. 15, No. 1, pp. 55-63.

Chang, A. S., & Leu, S. S. (2006). "Data mining model for identifying project profitability variables", *International Journal of Project Management*, Vol. 24, No. 3, pp. 199-206.

Cheng, M. Y., Tsai, H. C., & Liu, C. L. (2009). "Artificial intelligence approaches to achieve strategic control over project cash flows", *Automation in construction*, Vol. 18, No. 4, pp. 386-393.

Cheng, M. Y., Wu, Y. W., & Wu, C. F. (2010). "Project success prediction using an evolutionary support vector machine inference model", *Automation in Construction*, Vol. 19, No. 3, pp. 302-307.

Cheung, S. O., Wong, P. S. P., Fung, A. S., & Coffey, W. V. (2006). "Predicting project performance through neural networks", *International Journal of Project Management*, Vol. 24, No. 3, pp. 207-215.

Chi, S., Suk, S. J., Kang, Y., & Mulva, S. P. (2012). "Development of a data mining-based analysis framework for multi-attribute construction project information". *Advanced Engineering Informatics*, Vol. 26, No. 3, pp. 574-581.

Chua, D. K. H., Loh, P. K., Kog, Y. C., & Jaselskis, E. J. (1997). "Neural networks for construction project success", *Expert Systems with Applications*, Vol. 13, No. 4, pp. 317-328.

Cios, K. J., & Kurgan, L. A. (2005). "Trends in data mining and knowledge discovery", In *Advanced techniques in knowledge discovery and data mining* (pp. 1-26), Springer London.

CRISP-DM. (2003) Cross Industry Standard Process for Data Mining

Doloi, H. (2009). "Analysis of pre-qualification criteria in contractor selection and their impacts on project success", *Construction Management and Economics*, Vol. 27, No. 12, pp. 1245-1263.

Doloi, H., Sawhney, A., Iyer, K. C., & Rentala, S. (2012). "Analysing factors affecting delays in Indian construction projects". *International Journal of Project Management*, Vol. 30, No. 4, pp. 479-489.

Fallahnejad, M. H. (2013). "Delay causes in Iran gas pipeline projects". *International Journal of Project Management*, Vol. 31, No. 1, pp. 136-146.

Fan, H., AbouRizk, S., Kim, H., & Zaïane, O. (2008). "Assessing residual value of heavy construction equipment using predictive data mining model". *Journal of Computing in Civil Engineering*, Vol. 22, No. 3, pp. 181-191.

Faridi, A. S., & El-Sayegh, S. M. (2006). "Significant factors causing delay in the UAE construction industry",*Construction Management and Economics*, Vol. 24, No. 11, pp. 1167-1176.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). "Knowledge discovery in databases: An overview", *AI magazine*, Vol. 13, No. 3, pp. 57.

Frimpong, Y., Oluwoye, J., & Crawford, L. (2003). "Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study". *International Journal of project management*, Vol. 21, No. 5, pp. 321-326.

Han, S. H., Kim, D. Y., & Kim, H. (2007). "Predicting profit performance for selecting candidate international construction projects". *Journal of Construction Engineering and Management*, Vol. 133, No. 6, pp. 425-436.

Iranmanesh, S. H., & Mokhtari, Z. (2008). "Application of data mining tools to predicate completion time of a project", In *Proceeding of world academy of science, engineering and technology*. Vol. 32, No. 1, pp. 234-239.

Kaliba, C., Muya, M., & Mumba, K. (2009). "Cost escalation and schedule delays in road construction projects in Zambia". *International Journal of Project Management*, Vol. 27, No. 5, pp. 522-531.

Kaming, P. F., Olomolaiye, P. O., Holt, G. D., & Harris, F. C. (1997). "Factors influencing construction time and cost overruns on high-rise projects in Indonesia", *Construction Management & Economics*, Vol. 15, No. 1, pp. 83-94.

Kantardzic, M.(2011) *Data mining: concepts, models, methods, and algorithms*, John Wiley & Sons.

Kim, H., Soibelman, L., & Grobler, F. (2008). "Factor selection for delay analysis using Knowledge Discovery in Databases". *Automation in Construction*, Vol. 17, No. 5, pp. 550-560.

Ko, C. H., & Cheng, M. Y. (2007). "Dynamic prediction of project success using artificial intelligence". *Journal of construction engineering and management*, Vol. 133, No. 4, pp. 316-324.

Kohavi, R., & Sommerfield, D. (1995)"Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology", In *KDD* (pp. 192-197).

Marcoulides, G. A. (2005). "Discovering Knowledge in Data: an Introduction to Data Mining". *Journal of the American Statistical Association*, Vol. 100, No. 472, pp. 1465-1465.

Ma, Z., Lu, N., & Gu, W. (2008). "A decision support system for construction projects based on standardized exchanged documents", *Tsinghua Science & Technology*, Vol. 13, No. 1, pp. 354-361.

Mansfield, N. R., Ugwu, O. O., & Doran, T. (1994). "Causes of delay and cost overruns in Nigerian construction projects", *International Journal of Project Management*, Vol. 12, No. 4, pp. 254-260.

Munns, A. K., & Bjeirmi, B. F. (1996). "The role of project management in achieving project success". *International journal of project management*, Vol, 14, No. 2, pp. 81-87.

Odeh, A. M., & Battaineh, H. T. (2002). "Causes of construction delay: traditional contracts" *International journal of project management*, Vol. 20, No. 1, pp. 67-73.

Sambasivan, M., & Soon, Y. W. (2007). "Causes and effects of delays in Malaysian construction industry", *International Journal of project management*, Vol. 25, No. 5, pp. 517-526.

Shadrokh, S., & Aghdashi, S. (2012) "Data Mining in Construction's Project Time Management-Kayson Case Study"

Shahrabi, J., & Taghavi, Z. S. (2012) "A Data Mining Model For Feasibility Analysis Of Mineral Projects". *International Journal of Advances in Engineering & Technology*, Vol. 4, No. 2.

Sharma, S., & Osei-Bryson, K. M. (2009). "Framework for formal implementation of the business understanding phase of data mining projects", *Expert Systems with Applications*, Vol. 36, No. 2, pp. 4114-4124.

Son, H., Kim, C., & Kim, C. (2012). "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables", *Automation in Construction*, Vol. 27, No. 1. pp. 60-66.

Yang, J., & Honavar, V. (1998). "Feature subset selection using a genetic algorithm", In *Feature extraction, construction and selection* (pp. 117-136), Springer US.

Zack, J. G. (2003). "Schedule delay analysis; is there agreement?", In *Proc., PMI-CPM College of Performance Spring Conf* (pp. 7-9).