Automation in Construction 17 (2008) 550 – 560

AUTOMATION IN CONSTRUCTION

www.elsevier.com/locate/autcon

Factor selection for delay analysis using Knowledge Discovery in Databases

Hyunjoo Kim [a,*], Lucio Soibelman [b], Francois Grobler [a]

[a] *US Army Engineer Research and Development Center, IL, USA*
[b] *Carnegie Mellon University, PA, USA*

Accepted 3 October 2007

Abstract

Today's construction project has become a very complex, high-risk, multiparty endeavor. Construction projects are composed of many interrelated elements of labor, cost, material, schedule, and other resources, making it difficult to discern which factors were the main causes for delay on a given project. Were all relevant factors to be considered, it would become an overwhelming task. On the other hand, it would be very difficult to know on which factors to focus if only a limited number of factors were to be considered. This paper presents a methodology for factor selection; identifying which factors in an on-going construction project contribute most to the experienced delays. Factor selection is defined as the process of finding relevant factors among a large set of original attributes with the objective of best representing the original dataset and utilizes Knowledge Discovery in Databases (KDD), which is a data analysis process to discover useful knowledge in a large database. A specific construction project has been analyzed to identify main factors of construction delays through the process of statistical measurements and machine learning algorithms.

Keywords: Knowledge Discovery in Databases; Construction delay; Statistics; Machine learning; Factor selection

1. Motivation

In 2001 major US construction firms had over $200.8 billion dollars worth of contracts and since then this work load has been on the rise by about 8% annually. The construction industry has become a very complex, high-risk, multiparty business [1]. As the construction industry is growing, construction projects are also expanding in size and complexity. Identifying the main causes of delays in large construction projects is very difficult and often initiates disputes about responsibility for the delay. Many of these disputes end up in litigation because there are usually several parties and many factors to be involved.

Some of the causes of a construction delay are obvious in a project schedule, but others will be difficult to identify, due to uncertainty and many interdependent factors in the work environment. Construction projects are composed of many interrelated elements of labor, cost, material, schedule, and other resources, making it difficult to discern which factors were the main causes for delay on a given project. Considering too few elements in the delay analysis could result in incorrect conclusions and trying to consider all relevant factors is near impossible. Therefore it becomes important to know which factors to select and focus on. This paper presents a method for factor (or feature) selection.

Factor selection is defined as the process of finding relevant factors among a large set of original attributes with the objective of describing the original dataset. The method presented uses several methods from Knowledge Discovery in Databases (KDD) to identify and analyze the main causes of construction delays [2]. A specific project in the Resident Management System (RMS) has been analyzed to conduct case study. RMS is a large database used by the US Army Corps of Engineers to track construction contracts and progress information. In this paper we demonstrated how the main factors of construction delays were identified by applying the framework developed to support feature selection for KDD.

* Corresponding author. US Army Engineer Research and Development Center, PO Box 9005, Champaign, IL 61826. Tel.: +1 217 USA-CERL, ext. 7539; fax: +1 217 373 6724.
E-mail address: hyunjoo.kim@erdc.usace.army.mil (H. Kim).

0926-5805/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.autcon.2007.10.001

1.1. Requirements of factor selection in construction delays

Much research effort has been expanded on the subject of delay analysis. Some researchers worked on the classification of the delay [3] while others measured the impact of the delay in terms of lost productivity [4]. Also, the subject of compensable and non-compensable delays has been analyzed where the window method of analyzing delays has been evaluated [5]. However, little attention has been given to identifying the main cause(s) of the delays, likely due to the issues discussed in the previous section.

A systematic approach is necessary to understand what causes delays. Some projects may not experience schedule delays at all, even with certain problems (i.e., rainy weather, shortage of equipment, etc) while the same problems may result in serious delays in different work environments or with different project managers. The factor selection methodology must meet certain requirements in order to be successfully applied in construction delay analysis:

- Objective measurement is required: Traditionally, the causes of construction delay have been attributed to lack of leadership, teamwork, or training, which is very subjective [6]. Cloaked in such subjectivity it is hard to discover the real causes, and left unidentified and unresolved the same kind of delay problems may occur in future projects. This paper proposes an objective approach to detect the possible causes of project delays by analyzing construction project data collected on the project.
- Unbiased, systematic approach must include a suite of tools: Many research papers suggested how one would benefit from using a single tool or technique. However, it is hardly possible to find a single algorithm or tool that can be applied to today's entire construction projects — projects that are constantly changing in complex project environments, and with intricate multi-organization teams. This paper suggests that one would analyze the construction data more efficiently by applying several different factor selection tools and comparing and combining different techniques. Results of different factor selection techniques were measured and compared in this paper. There have been many efforts mainly by statistical and machine learning communities in identifying important factors (or attributes) from a dataset by measuring the relevance between input and output factors. The results indicate the importance of incorporating different approaches by compensating each other.
- Consider interrelations among factors: Construction data are composed of many interrelated elements of data such as cost, schedule, and resources such as labor, and materials. The factor selection method must identify the best set of factors (or feature) among numerous relevant factors in the large set of original attributes, while dealing with their complex relationships in the dataset.

1.2. Previous research on KDD applications

Factor selection draws from several tools in Knowledge Discovery in Databases (KDD). KDD can be considered an interdisciplinary field involving different concepts from machine learning, statistics, database query, and visualization. While the purpose of database technologies is to find efficient ways of storing, retrieving, and manipulating data, the main concern of the machine learning and statistical communities is to develop techniques for extracting knowledge from data [2].

Fig. 1 shows many different KDD applications in civil engineering where the left side of the graph represents the research on data preparation and the right side shows different types of KDD applications. Data preparation is an important area and many researchers worked on the preparation process

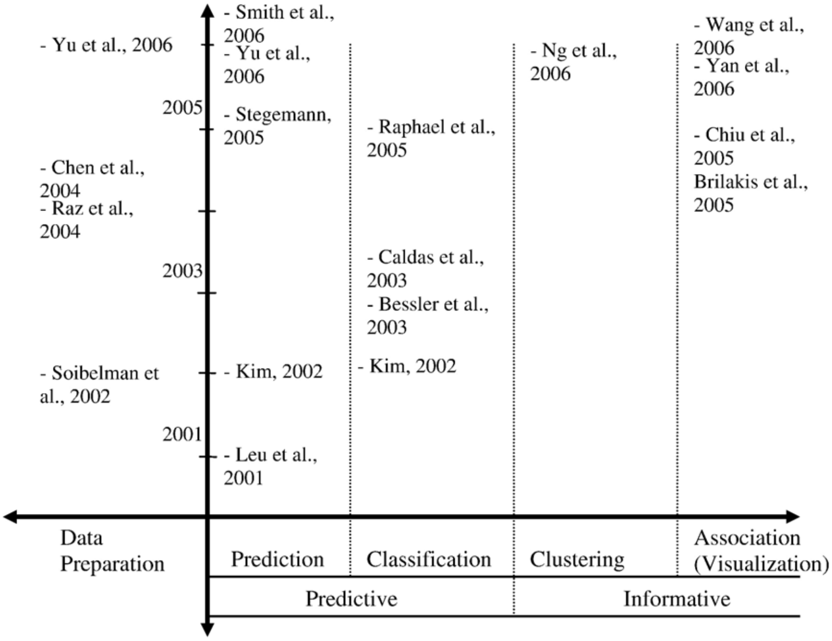

Fig. 1. Comparison of KDD application in construction.

due to the fact that construction data usually consist of incomplete or inaccurate data. Thus, various techniques were introduced to improve quality of data [2,7–9].

As for data analysis, there are two different types of patterns to be found in KDD process: predictive (prediction and classification) and informative (clustering and association). It is interesting to note in Fig. 1 that in early 2000s, the applications of data analysis were limited to predictive applications. However, the applications are being extended to diverse areas such as clustering and association (visualization) that require heavy computational power. Predictive patterns are built to solve a specific problem of predicting one or more attributes in a database. On the other hand, the informative patterns are not utilized to solve a specific problem, but rather to present interesting patterns that a domain expert might not already know.

- Prediction refers to discovering predictive patterns where the attribute being predicted is real-valued. Predictive patterns are important not only in how well they predict the desired unknown quantity, but also, in their ability to convey the interesting patterns to the domain expert. Smith and Saitta [10] used KDD algorithm to estimate model populations and Yu et al. [11] predicted cost estimation in conceptual design. Kim [12] and Leu et al. [13] applied neural network as a KDD algorithm for prediction.
- Classification is similar to a prediction pattern, except that it predicts the value of a nominal or categorical attribute instead of real-valued attribute. The predicted attribute defines a class in classification. Raphael et al. [14] used decision tree and correlation measurement for feature extraction. Caldas and Soibelman [15] applied classification technique for construction documents that requires much of preparation process.
- Clustering is also called a segmentation model. The goal of clustering is to take a set of entities represented as records in a database and to partition them into a number of groups or clusters so that the entities within a cluster are similar. Ng et al. [16] proposed facility condition assessment using text clustering.
- Association is an informative pattern of the form $X \rightarrow Y$, where X and Y are statements about the values of attributes of a record. Association rules discovery algorithms search the space of all possible rules $X \rightarrow Y$ where X and Y are sets of items. Yan et al. [17] proposed video-derived information analysis for structural analysis and Wang and Ghosn [18] used Genetic Algorithm to explore relations among random variables for controlling the safety of a structural system. For visualization, Pande and Abdel-Aty [19] applied the KDD technique of visualization for real-time crash risk assessment and Brilakis et al. [20] developed a system for construction image indexing and retrieval.

Research review revealed that in many applications, a single tool or algorithm has been developed and applied to identify useful patterns. However, Fayyad and Uthrusamy [21] proposed that one would analyze data more efficiently by combining different tools and comparing different techniques. Construction data consist of many interrelated elements. The nature of construction project is dynamic and complex. Therefore, developing a single algorithm that can be applied to all different projects is not a realistic approach. Thus, an integrated approach for finding important patterns among construction databases by incorporating different algorithms and analysis tools is necessary.

1.3. Integrated approach

With the current development of numerous data analysis tools, it is impossible to say that any specific technique is better than another over all problems [34]. In considering the previous research, we recognize that data analysis may be improved by combining statistics with machine learning and comparing different results to correctly identify the main problems or factors for a particular situation in a project. This research utilized the approaches of statistical and machine learning techniques mainly for the following reasons:

- *Statistical approach*: There have been many efforts in converting the large amounts of data into useful patters or trends [22–24]. In addition to these new techniques, statistics is essential for the handling of uncertainty. On the other hand, most statistics methods have difficulties with computational complexity issues or dealing with a large amount of parameter values.
- *Machine learning approach*: As for machine learning, there has been an increased interest in learning systems because new formal methods and new techniques of implementation have been developed and both the cost and speed of running learning systems have improved dramatically [18]. Machine learning systems are computer programs that automatically improve their performance through experience using tools which include inductive inference of decision trees, neural network learning, genetic algorithms, explanation based learning, and reinforcement learning and so on. Even though machine leaning techniques usually yield relatively high accuracies compared to statistical tools, Melhem et al. [25] suggested that machine learning algorithm might result in misclassification or misleading.Therefore, this paper describes a procedure in the next section to maximize the benefits of each technique by focusing on the interplay between statistical measurements and machine learning algorithms.

1.4. Application to construction delay

For the purposes of this paper construction delay was measured by the difference between planned schedule and actual schedule for a construction activity. The difference between planned/actual schedules was considered along with existing float, based on the fact that the delay would not make an impact on the following activities as long as the activity was completed within the positive float remaining. Project documentation in daily, weekly or monthly written report should be prepared to evaluate the schedule delay. As the construction projects become bigger in size, it is common to find a large amount of databases storing daily conflicts or problems in jobsites. The main cause (or attribute) of delay in this research is referred to as the most frequent problems occurred in a project, causing the project to be delayed in schedule. A common approach to identify the main causes of delay(s) is classify the different activities into similar ones so that a similarity (or useful pattern) of delay may be found during the classification through data analysis. A pattern is an expression of describing facts in a subset of a set of facts [2]. The expression is called a pattern if it is simpler than the enumeration of all facts in the subset of facts. For example, "If the cost of any activity is less than a certain threshold then historical data proves that there is a high probability that the activity is on schedule" would be one such pattern for an appropriate choice for construction cost.

The proposed methodology for factor selection is shown in Fig. 2. As the first step of data analysis process, frequency charts were used for data exploring. Frequency chart can often identify effects (both expected and unexpected) in the dataset very quickly and easily providing a visual impression of the data distribution. As the second step, the factor selection process was conducted with two approaches such as statistical and machine learning approaches. Statistical approaches can be grouped into three methods such as correlation matrix, factor analysis, and Bayesian networks. The machine learning approach utilizes an inductive learning algorithm called wrapper approach. Finally, the machine learning approach was applied to select the final set of most important attributes. In the verification process, the results of the statistical and machine learning approaches were compared and combined to compensate for each other's limitations.

2. Case study

In order to test the feasibility of the factor selection approach proposed in this paper, a case study was conducted with the data from a flood control project in Fort Wayne, Indiana with the dataset from their Resident Management System (RMS) provided by US ARMY Corps of Engineers. This research analyzed a relational database called the RMS where resident engineers in the Corps of Engineers store daily, weekly or monthly reports to evaluate construction projects and to keep tract of on-going activities and differences between original and actual durations. Daily activities are automatically updated in RMS that a project manager may compare between the as-planned and the as-built schedule and closely monitor which instances of suspected delay become apparent. The specific goal of the case study is to identify which factors in an on-going construction project contribute most to the significant delay in a project so that construction personnel may resolve the problems correctly during construction process.

2.1. Characteristics of data

The purpose of RMS is to manage construction projects, to perform program control and to provide the capability to exchange (design, scheduling and construction) data with contractors through the use of specialized modules. RMS stores data from previous construction projects data that included construction project planning, contract administration, quality assurance, payments, correspondence, submittal management, safety and accident administration, modification processing, and management reporting (Fig. 3).

Table 1 shows variable names and descriptions of RMS database. It is PC-based, LAN-compatible. RMS also has automated, single-entry data exchange/communication capabilities with Corps-wide systems (CEFMS, PROMIS, SAACONS and so on.). The construction project used in our case study was

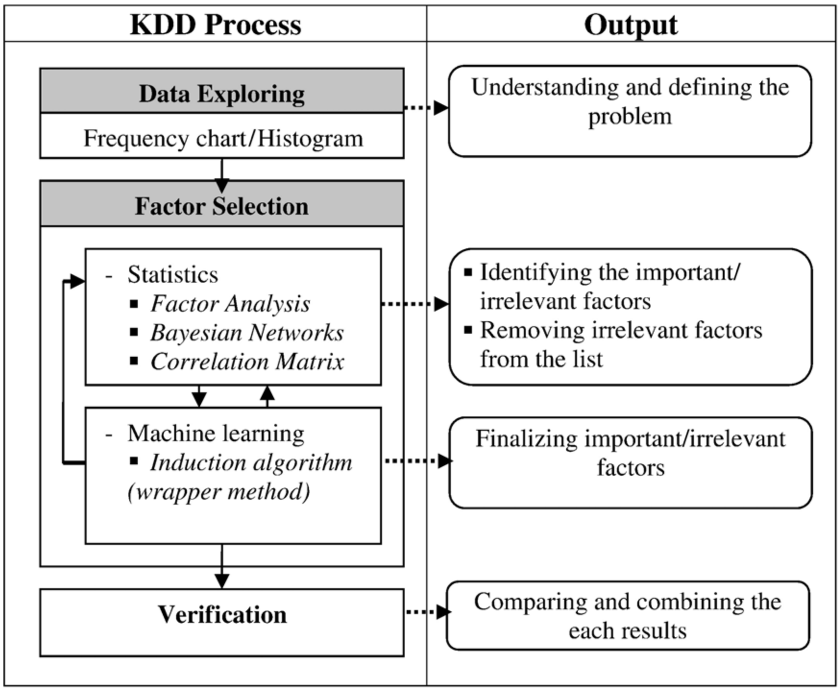
KDD Process
Output
Data Exploring
Frequency chart/Histogram
Factor Selection
- Statistics
▪ Factor Analysis
▪ Bayesian Networks
▪ Correlation Matrix
- Machine learning
▪ Induction algorithm (wrapper method)
Verification
Understanding and defining the problem
▪ Identifying the important/ irrelevant factors
▪ Removing irrelevant factors from the list
Finalizing important/irrelevant factors
Comparing and combining the each results

Fig. 2. Main procedure of factor selection.

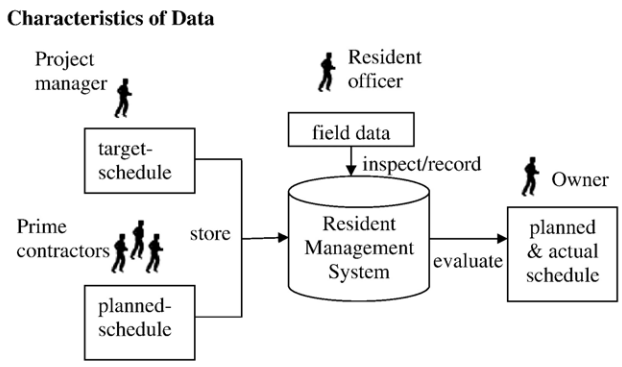

Fig. 3. Transactions of the RMS in a construction project.

designed to limit the damage caused by frequent flooding that has occurred on average every four to five years. Three construction phases (Phase I: CTRL-EAST, $4,488,450.21, Phase II: East-North, $12,107,880.46, and Phase III: CTRL, $ 6,018,981.54) constituted the Fort Wayne Flood Control Project to provide enhanced flood protection to a large part of the central area of the City of Fort Wayne, Indiana. The initial data survey from data acquired on phases I and II demonstrated that the project was behind schedule on 54% of its activities of installing drainage pipelines.

3. Statistical approach

The statistical analyses such as frequency charts, correlation matrix, factor analysis, and Bayesian networks for this research were performed as follows.

3.1. Frequency charts

Frequency charts represent the simplest method for analyzing data. They are often used as one of the exploratory procedures to review how different values are distributed in the data. This simple statistical tool can be used to show distributions of data related to delayed or non delayed activities. It can reveal the number of delayed and non delayed instances in terms of seasons (spring, summer, fall, or winter), the number of instances from different locations, and so on.

The number of occurrences can also be summarized with means, median and standard deviations. Thus, one may tabulate the frequency of different causes leading to construction delays during a project (e.g., which activity or factors are most frequent for the construction delays?). In general, if a data set includes categorical data, one of the first steps for data analysis is to draw charts for those categorical variables. Frequency charts can often identify effects (both expected and unexpected) in the data very quickly and easily. Fig. 4 enables instant evaluation and visualization of the frequencies for each factor and provides an accurate visualization of the data distribution for more than one factor. The initial data survey for the case study demonstrated that one activity called Installation of drainage pipelines" for the tasks such as excavating the ground, installing pipelines, backfilling compacted, and erosion protection (detailed factors are shown in Table 1) was behind schedule 54% of the time (Table 1 shows the main attributes of the sub-activities).

3.2. Correlation

3.2.1. Algorithm

In probability theory and statistics, correlation, also called correlation coefficient, is a numeric measure of the strength of linear relationship between two random variables. Correlation is a single number that describes the degree of relationship between two variables.

Table 1
RMS attributes for construction delay

Column no.	Variable name	Description
1	Durat	Duration; no. of duration for an activity
2	Tot_Float	Total float; no. of total float for an activity
3	Locat	Location;
4	Const_phase	Construction phase; A = phase 1, B = phase 2, C = phase 3, D = phase 4
5	Work_per	Working percentage; percentage of current work
6	Inc_Draw	Incomplete drawing; 0 = incomplete, 1 = complete
7	Shifts	0 = day, 1 = night
8	Reg_Change	Regulatory changes; 0 = no, 1 = yes
9	Perm_Approval	Permit approval; 0 = required, 1 = not required
10	Des_Mod	Design modification; 0 = no, 1 = yes
11	Re_work	0 = no, 1 = yes
12	Envir_Issues	Environmental issues; 0 = no, 1 = yes
13	Const_Meth	Construction method; 0 = surface, 1 = near-surface, 2 = mid-depth, 3 = off-bottom, 4 = bottom tow methods
14	Rain/Snow	1 = rain, 2 = snow, 3 = no rain/snow
15	Wind	Wind blow, MPH
16	Act_Desc	Activity description
17	Temp	No of temperature
18	Cha_Order	Change order; 0 = no, 1 = yes
19	EquipHrs	Equipment hours, no of equipment hours
20	Equip_Break	Equipment breakdown; hours of equipment breakdown
21	Impr_Equip	Improper equipment
22	Dam_Good	Damaged goods
23	Impr_Tool	Improper tools, 0 = no, 1 = yes
24	Mat_Dly	Material delivery; 0 = no problem with delivery, 1 = yes
25	Inacc_Est	Inaccurate estimates; 0 = no, 1 = yes
26	Subcon_Interfer	Subcontractor interference; 0 = no, 1 = yes
27	Subcon_Delay	Subcontractor delay; 0 = no, 1 = yes
28	Equip_no	Equipment number, no of equipment used
29	Short_Equip	Shortage of equipment, 0 = no, 1 = yes
30	MAX_Tem	MAX temperature
31	MIN_Tem	MIN temperature
32	Season	Season, 0 = spring, 1 = summer, 2 = fall, 3 = winter
33	Incom_S_Survey	Incomplete site survey; 0 = no, 1 = yes
34	Labor_No	Labor number
35	Labor_Hrs	Labor hours
36	Safe_Viol	Safety violation; 0 = no, 1 = yes
37	Inspect_ID	Number of inspection a day
38	Rain_End	Time for rain/snow to end
39	Weekends	Work in weekends, 0 = no, 1 = yes
40	Crew_Size	No of crews
41	Type_Soil	Type of soil; 0 = sandy, 1 = loamy, 2 = mixed
42	Tren_Bot	Trench bottom; 0 = flat, 1 = hard, 2 = rocky

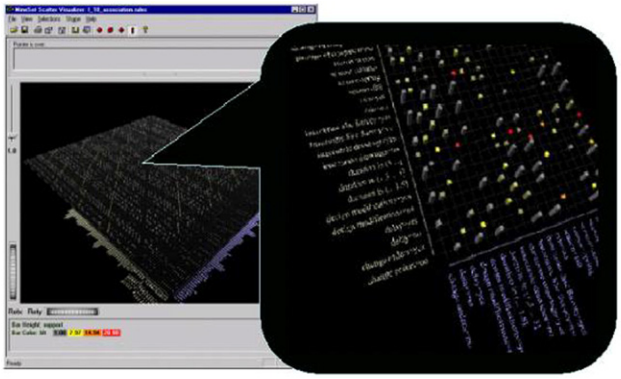

Fig. 4. Frequency chart of RMS data (Mineset, SGI).

Correlation matrix is one of the most common and most useful statistical tools where the data analyst can obtain useful insights on important factors being able to identify some causal relationships in construction delay data. Two triangles (values below and to the left of the diagonal-lower triangle and above and to the right of the diagonal-upper triangle) of a correlation matrix are always mirror images of each other. That is because the correlation of variable x with variable y is always equal to the correlation of variable Y with variable X.

3.2.2. Result

By applying the correlation matrix approach "Temperature" was identified as the most correlated factor. Fig. 5 shows that the important factors according to the correlation analysis were "Temperature", "Material Delivery", "Weekend", "Inaccurate Site Survey", "Subcontractor Delay", "Inaccurate Drawing", "Safety", "Weather (Rain/Snow)", and "Seasons".

3.3. Bayesian networks

3.3.1. Algorithm

Bayesian networks (also known as causal or probabilistic networks) are currently one of the most popular uncertainty knowledge representation techniques. A Bayesian network or belief network is a directed acyclic graph of nodes representing variables and arcs representing dependence relations among the variables. If there is an arc from node A to another node B, then A is a *parent* of B. If a node has a known value, it is said to be an *evidence* node. A node can represent any kind of variable, be it an observed measurement, a parameter, a latent variable, or a hypothesis [26]. Nodes are not restricted to representing random variables. A Bayesian network is a representation of the joint distribution over all the variables represented by nodes in the graph.

3.3.2. Result

In the case study, Bayesian network demonstrated that the probability of delays was directly affected by the "Inaccurate Site Survey", "Shortage of Equipment", "No. of Workers", "Season", and "Design Modification". The model also revealed that "Rain/Snow" had some effect on the delay, while "Inaccurate Estimation" and "Change Order" may have caused the delay via other attributes. On the other hand, Bayesian networks suggested that the most efficient way to avoid delays was to target "Inaccurate Site Survey" and the "Shortage of Equipment". Although there were causal networks among attributes such as "Inaccurate Site Survey", "Shortage of Equipment", "Inaccurate Estimation", "Season", and "Duration", "Incomplete drawing", " Temperature" and "Weekends", some networks in Bayesian Networks were not clear, requiring an expert to add information to allow that all the relationships in the graph could be understood.

3.4. Factor analysis

3.4.1. Algorithm

Factor analysis is a mathematical tool that can be used to examine a wide range of data sets. It has been used in disciplines as diverse as chemistry, sociology, economics, and psychology among many others. The purpose of factor analysis is to discovery simple patterns in the relationships among the variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a much smaller number of variables called factors. Factor analysis can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (or factors). The goal of this

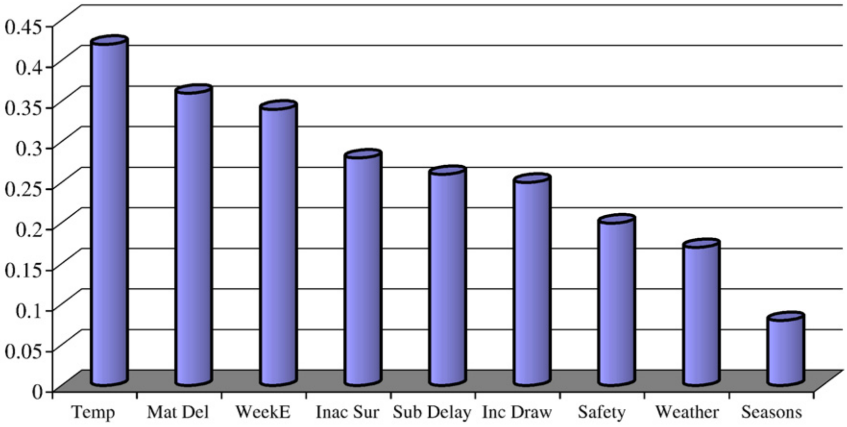

Fig. 5. Results from correlation matrix from Fort Wayne project.

tool is to find a way of condensing the information contained in a number of original variables into a smaller set of dimensions (factors) with a minimum loss of information [27]. The rating given to any one attribute is partially the result of the influence of other attributes. The statistical algorithm deconstructs the rating (called a raw score) into its various components, and reconstructs the partial scores into underlying factor scores.

3.4.2. Result

Factor analysis for the project in Fort Wayne resulted in the selection of the following 23 different attributes: Inaccurate Site Survey, Change Order, Equipment Breakdown, Weekends, Work Percentage, Regulatory Change, Work Permit Approval, Design Modification, Equipment Hours, Material Delivery, Season, Temperature Max, Temperature Min, Original Duration, Inaccurate Estimation, Improper Tools, Damaged Goods, Improper Equipment, Location, Total Float, Construction Method, Temperature, Safety.

3.5. Summary of statistical approaches

Different statistical approaches were compared in Table 2 to explore the dataset. In Table 2, the results of each method are diversely spread. One advantage of conducting different approaches is to increase the probability of not selecting biased factors. With statistical techniques combined, there are 27 factors chosen by each technique while 15 factors were found to be non-relevant. Among the 15 factors chosen to be non-relevant, we examined carefully and found that two factors (Activity_ID, and Inspection_Time) were data in constant types that mostly contain descriptions and do not deliver any significance to important factors. Next, we discovered that seven factors such as "Wind", "Subcon_Interfere", "Labor_No", "Type_Soil", "Equip_NO", "Environmental_Issues" and "Shifts" contain monotonic variables where most values are almost the same except for a few instances. It is understood in this research that monotonic values of the dataset could occur when data were recorded by erroneous input or the jobsite situation did not change during the construction. It is also noted that all the statistical tools identified four common factors ("Inc_Draw", "Temp", "Season", and "Weekends") to be relevant to the construction delay.

4. Machine learning approach

In supervised classification learning, an induction algorithm is used to form a classifier to predict a class label for an unseen instance through training on a set of instances with class labels. This research utilized an induction algorithm in identifying the important factors. The example of identifying relevance between factors (or features) is as follows: Typically the feature subset which performs best for the induction algorithm will be selected. Machine learning method considers the attribute X_i relevant if knowing its value can change the estimates for Y. In other words, an attribute X_i is relevant if the probability of the output given all attributes can change when we eliminate knowledge about the value of X_i. An attribute X_i is relevant if and only if every expression there exists some x and y such that $p(Y|X)$ contains X_i [28].

4.1. Induction learning algorithm

Decision tree (inductive learning) algorithm was used in this work to identify important factors. Decision tree is a inductive learning algorithms and has been applied to many different applications as a simple and yet successful form of a machine learning algorithm. During the last decade, there was much research on decision tree induction [29,30].

Table 2
Comparison of different factor selections

Variable name	Correlation matrix	Bayesian networks	Factor analysis	Variable name	Correlation matrix	Bayesian networks	Factor analysis
Durat		●	●	Inacc_Est		●	●
Tot_Float			●	Subcon_Interfer			
Locat			●	Subcon_Delay	●		
Const_phase				Equip_no			
Work_per			●	Short_Equip		●	
Inc_Draw	●	●	●	MAX_Tem			●
Shifts				MIN_Tem			●
Reg_Change			●	Season	●	●	●
Perm_Approval			●	Inacc_S_Survey	●	●	●
Des_Mod		●	●	Labor_No	●		
Re_work				Labor_Hrs			
Envir_Issues				Safe_Viol			●
Const_Method			●	Inspect_ID			
Rain/Snow	●	●		Rain_End			
Wind				Weekends	●	●	●
Act_Desc				Crew_Size			
Temp	●	●	●	Type_Soil			
Cha_Order		●	●	Impr_Equip			●
EquipHrs			●	Dam_Good			●
Equip_Break			●	Impr_Tool		●	●
Mat_Dly	●		●	Act_ID			

Table 3
Factor selection by induction learning algorithm

Variable name	Description
Durat	Duration; no. of duration for an activity
Inc_Draw	Incomplete drawing; 0 = incomplete, 1 = complete
Incom_S_Survey	Incomplete site survey; 0 = no, 1 = yes
Short_Equip	Shortage of equipment, 0 = no, 1 = yes
Weekends	Work in weekends, 0 = no, 1 = yes
Inspection_Time	Time. of inspection during a day
Crew_Size	NO. of crews
Season	Season, 0 = spring, 1 = summer, 2 = fall, 3 = winter
Rain/snow	1 = rain, 2 = snow, 3 = no rain/snow

Decision trees use entropy criteria for splitting nodes. Given a node t, the splitting criterion used is Entropy$(t)=\sum_i -p_i \log p_i$ [31] where P_i is the probability of class i within node t. An attribute and split are selected that minimize entropy. Splitting a node produces two or more direct descendants. Each child has a measure of entropy. The sum of each child's entropy is weighted by its percentage of the parent's cases in computing the final weighted entropy used to decide the best split. If there are n equally probable possible messages, then the probability p of each is $1/n$ and the information conveyed by a message is $-\log(p)=\log(n)$.

In general, if we are given a probability distribution $P=(p1, p2,\ldots pn)$ then the Information conveyed by this distribution, also called the entropy of P, is:

$$I(P) = -(p1^* \log(p1) + p2^* \log(p2) + \ldots\ldots + pn^* \log(pn)) \tag{1}$$

[31] Gain(X,T) defined as

$$\mathrm{Gain}(X,T) = \mathrm{Info}(T) - \mathrm{Info}(X,T) \tag{2}$$

[31].

This represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been obtained, that is, this is the gain in information due to attribute X.

4.2. Results of inductive learning algorithm

Using the inductive learning algorithm of C4.5 decision tree, the results of decision tree (induction learning) method was measured in error rates. The common error rate can be defined as the ratio of the number of errors to the number of cases examined.

$$\text{Error rate} = \text{number of error}/\text{number of cases}$$

[32].

Holding out a number of randomly selected samples as a test data set and not using these cases at all as part of the training process allows these samples to be used to obtain a good estimate of the true error rate [32]. 10-fold cross validation was utilized to improve over the holdout method in this research. The data set was divided into 10 subsets, and the holdout method was repeated 10 times. Each time, one of the 10 subsets was used as the test set and the other subsets $(10-1=9)$ were put together to form a training set. Then the average error across all 10 trials was computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set 9 $(=10-1)$ times. The variance of the resulting estimate is reduced as the number of the training sets is increased.

Using the induction learning approach (or wrapper approach) in the case study, nine different factors were selected as shown in Table 3. In this approach, the error rate of the induction decision tree algorithm (C4.5) was measured at 15.2±2.3% (confidence level: 5%). Upon reviewing the results chosen by inductive learning algorithm, it is noted that one factor of "Inspection_Time" was identified as irrelevant since it contains the constant values and doesn't deliver any significant value. Also, one of the common factors, "Temperature" which was chosen to be relevant by all the statistical tools was not selected in the perspective of

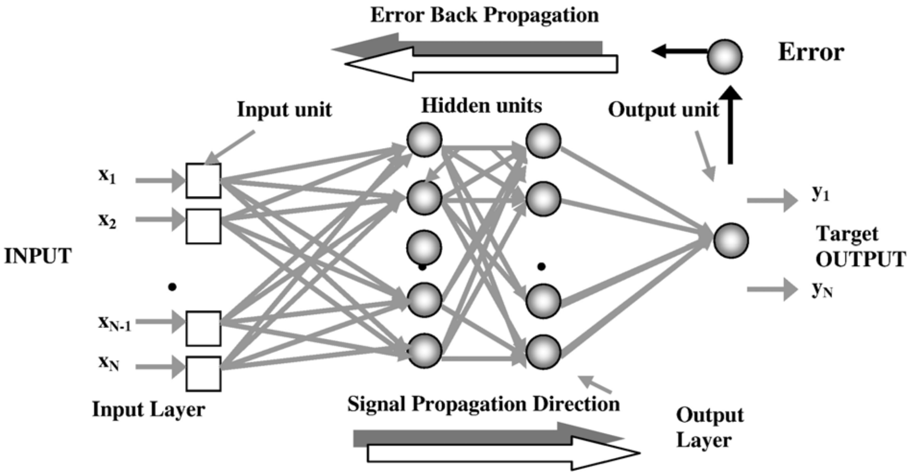

Fig. 6. Schematic diagram of back-propagation neural networks with two hidden layers.

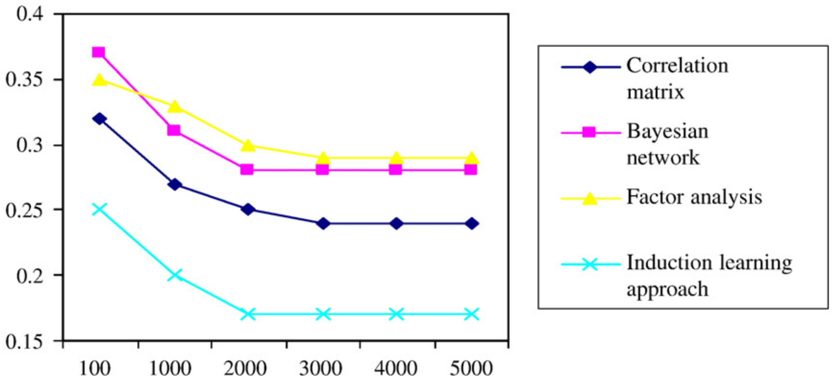

Fig. 7. Error rates of neural networks.

inductive machine learning. So, we concluded that there are some discrepancies of factor selection between statistical and machine learning algorithms. These discrepancies were further considered to revise the factor selection process in the validation process.

5. Comparison of different factor selections

Comparison was conducted through the implementation of the back-propagation neural network which is currently the most popular artificial neural network [33]. Its flexible structure is composed of several arrays of weight matrices. These weight matrices represent layers of small information processing nodes which perform calculate simple nonlinear functions of their inputs and pass the result on to the neurons in next layer. This modeling algorithm adjusts a large number of numerical parameters to minimize model error.

The neural networks is run to learn from the existing dataset through the generalization of the network. Fig. 6 shows the schematic diagram of back-propagation neural network that was developed during the comparison process where each factor chosen in statistic and machine learning approaches was set as inputs to predict the activity duration. Types of each variable were converted to numbers as input values. The output value was the delay for the activity of installing drainage pipeline. The training cycle was repeated for each case with small adjustments being made in the weights after each case. In deciding the appropriate number of hidden layers and the best learning rate, a great number of neural networks was run. In this case study, it was found that the best result was given with 1% learning rate and a 4 layers back-propagation neural networks architecture. The comparison of error rates was measured when different attribute selection tools like factor analysis, Bayesian networks, correlation, and inductive algorithm of wrapper approach were compared as the number of cycle increases. Fig. 7 shows the optimum point around 3000 cycles where the error rate stops decreasing with four different methods (*X* axis represents the number of cycles of neural networks of each approach while *Y* axis shows error rates of each tool in percentage). The result of neural networks shows that the wrapper method had the lowest error rate of 17% and factor analysis had the highest error rate of 29%.

Even though the machine learning algorithm produced the factor selection with the lowest error rate, it was found that it included one factor that had no significant meaning. Our domain knowledge found that a factor “Inspection_Time” should be irrelevant factor which was confirmed by statistical tools. So, out of nine factors chosen by wrapper method, “Inspection_Time” was removed from the whole list. Also, “Temperature” was shown to be relevant by all the statistical approaches. However, the induction learning method didn't find the factor to be relevant. Based on the domain knowledge that most construction operations are effected by hot or cold weather, temperature was included as an important factor. When we closely examined the reason that the factor was not measured to be relevant, we found that temperature in the data was of numeric values (−32°F~102°F). Therefore, we adjusted the range of numeric values in temperature by converting the numeric values into five nominal (categorical) values such as extreme_hot, hot, normal, cold, and extreme_cold. This attempt turned out to be quite successful and eventually increased the accuracy of wrapper method to 12.4% as shown in Fig. 8.

5.1. Problem encountered

From the case study conducted in the research, this section describes the lessons learned while applying Knowledge Discovery in Databases (KDD) to identifying the causes of construction delays.

- In order to prepare the data, much effort was necessary to improve the quality of data. Most common problems that were encountered during the case study were categorical

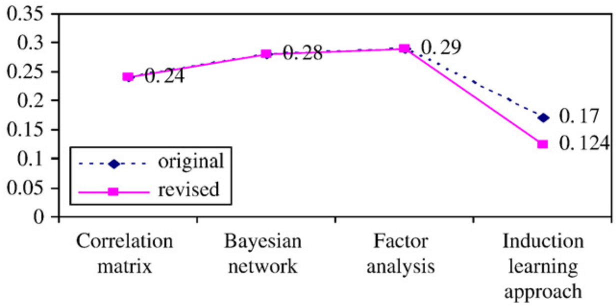

Fig. 8. Final error rates of neural networks.

data, missing or empty data, monotonous data and too low dimensionality (the number of variables to be considered). Thus, the proper way of handling and enhancing those values are desired for better data analysis. More description on data preparation for construction KDD can be found in Soibelman and Kim [2].

- Even though machine learning usually yielded higher accuracies in terms of data analysis, there still existed chances of misleading results or misclassification. Therefore, automatic factor selection while relying on machine learning only may result in mistakes.
- On the other hand, statistical community has been traditionally doing research on factor selection. So, there are many different methods to identifying important factors in statistics. This research implemented some of the popular methods such as frequency charts, correlation matrix, factor analysis, and Bayesian networks. One of the advantages of statistical approaches was that data was processed very quickly compared to running machine learning tools. Thus, statistical data selection is a good choice where frequent data analysis is required.
- Some research papers suggested that one would benefit from using a single tool or algorithm. However, we concluded that construction data could be analyzed more efficiently and accurately by maximizing the strong points of each technology. This was accomplished by focusing on the comparison between statistical and machine learning algorithms in an integrated approach.
- Various methods have been applied to identify important factors. It was found that each algorithm would produce similar but not exactly the same results. Therefore, the question of which algorithm works best in a particular situation is to be answered. In order to find the best algorithm for a specific application, further research is needed.
- In this paper, data analysis was conducted as KDD process to identify the main causes of schedule delays. However, its possible applications can be extended to identifying the causes of various applications such as cost overrun and quality assurance and so on. By applying the KDD process to different construction areas, the construction personnel could have a better way of managing his/her construction project.

6. Conclusions

As the construction industry is adapting to the new computer technologies in terms of hardware and software, computerized construction data are becoming more and more available, However, the our reality is that most data in construction projects is used only for communication purposes and stored in a file or a database without being analyzed. This research presented a framework by applying KDD technology to identifying the main factors (or causes) for construction delays.

Factor selection is an important issue for data analysis. It is used to select the most relevant factors from the data. By selecting only the relevant features of the data, higher predictive accuracy can be achieved and the computational load for the classification or prediction can be reduced. There have been many efforts mainly by statistical and machine learning communities in converting the large amounts of data into useful patterns or trends. Many research papers suggested how one would benefit from using a single tool or technique. However, this research found that one would analyze the construction data more efficiently by combining different techniques, statistical and machine learning techniques. Results of different factor selections using back-propagation neural networks were measured and compared. The results indicate the importance of incorporating different approaches by compensating each other.

Even though new research papers claim that new tools and techniques (machine learning tools, pattern recognitions and so on) delivered good performances in many applications, it is important to consider that the majority of new machine learning and pattern recognition tools were developed based on statistical principles. The main problem with learning algorithms is that they often have to face a large number of features so that accuracies of some classifiers become low.

To avoid subjective judgment on identifying the main factors for delays, a systematic approach with knowledge discovery process was utilized in this research where different algorithms were compared for unbiased data analysis and interrelations among factors were sought between many different factors in RMS database.

The purpose of this paper is to develop knowledge discovery procedure in the area of identifying the main factors of delays so that construction personnel may learn from on-going projects and apply the learned information to their current or future projects. In this paper, the authors introduced a various approach to identify important factors of construction delays. Even though machine learning is at the core of data analysis algorithms, machine learning algorithm might result in incorrect classification or prediction. By recognizing that machine learning may be improved by combining statistics with machine learning, this paper described a procedure to handle the limitations of each technique by integrating and combining the two techniques.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant number 0093841 (CAREER). The authors would like to thank the U.S. Army Corps of Engineers for providing us access to their RMS database.

References

[1] C. Semple, F.T. Hartman, G. Jergeas, Construction claims and disputes: causes and cost/time overruns, Journal of Construction of Engineering and Management, ASCE 120 (4) (1994) 785–795.

[2] L. Soibelman, H. Kim, Data preparation process for construction knowledge generation through knowledge discovery in databases, Journal of Computing in Civil Engineering, ASCE 16 (1) (2002) 39–48.

[3] S. Kartam, Generic methodology for analyzing delay claims, Journal of Construction Engineering and Management, ASCE 125 (6) (1999) 401–419.

[4] H. Lee, H. Ryu, J. Yu, J. Kim, Method for calculating schedule delay considering lost productivity, Journal of Construction Engineering and Management, ASCE 131 (11) (2005) 1147–1154.

[5] M.R. Finke, Window Analyses of compensable delays, Journal of Construction Engineering and Management, ASCE 125 (2) (1999) 96–100.

[6] D. Arditi, H.M. Gunaydin, Factors that affect process quality in the life cycle of building projects, Journal of Construction Engineering and Management, ASCE 124 (3) (1998) 194–203.

[7] O. Raz, R.B. Buchheit, M. Shaw, P. Koopman, C. Faloutsos, Detecting semantic anomalies in truck weigh-in-motion traffic data using data mining, Journal of Computing in Civil Engineering, ASCE 18 (4) (2004) 291–300.

[8] P. Chen, R.B. Buchheit, J.H. Garrett, S. McNeil, Web-vacuum: web-based environment for automated assessment of civil infrastructure data, Journal of Computing in Civil Engineering, ASCE 19 (2) (2005) 137–147.

[9] W. Yu, H. Lin, A VaFALCON neuron-fuzzy system for mining of incomplete construction databases, Automation in Construction 15 (1) (2006) 20–32.

[10] I. Smith, S. Saitta, Multiple-model updating to improve knowledge of structural system behavior, 17th Analysis and Computation Specialty Conference, ASCE, 2006, pp. 1–10.

[11] W. Yu, C. Lai, W. Lee, A WICE approach to real-time construction cost estimation, Automation in Construction 15 (1) (2006) 12–19.

[12] H. Kim, Knowledge discovery and machine learning in a construction project database. Ph.D. thesis, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 2002.

[13] S. Leu, C. Chen, S. Chang, Data mining for tunnel support stability: neural network approach, Automation in Construction 10 (4) (2001) 429–441.

[14] S. Saitta, B. Raphael, I. Smith, Supporting Engineers during system identification, Proceedings of Computing in Civil Engineering, ASCE, 2005, pp. 324–332.

[15] C. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Automation in Construction 12 (4) (2003) 395–406.

[16] H. Ng, A. Toukourou, L. Soibelman, Knowledge discovery in a facility condition assessment database using text clustering, Journal of Infrastructure Systems, ASCE 12 (1) (2006).

[17] A. Yan, M. Fraser, J. Lu, Large scale simulation and data analysis, Proceedings of Computing in Civil Engineering, ASCE, 2005, pp. 203–210.

[18] J. Wang, M. Ghosn, Hybrid data mining/genetic shredding algorithm for reliability assessment of structural systems, Journal of Structural Engineering, ASCE 132 (9) (2006).

[19] A. Pande, M. Abdel-Aty, Application of data mining techniques for real-time crash risk assessment on freeways, Applications of Advanced Technology in Transportation, ASCE (2005) 250–256.

[20] I. Brilakis, L. Soibleman, Y. Shinagawa, Material-based construction site image retrieval, Journal of Computing in Civil Engineering, ASCE 19 (4) (October 2005).

[21] U. Fayyad, R. Uthrusamy, KDD process for extracting useful knowledge from volumes of data, Proceedings of KDD-94: the AAAI-94 workshop on Knowledge Discovery in Databases, AAAI Press report, 1994.

[22] F. Sadek, E. Simin, Peak non-Gaussian wind effects for database-assisted low-rise building design, Journal of Engineering mechanics, ASCE 128 (5) (2002).

[23] H. Yi, T. Mulinazzi, J. Lee, Traffic-count based distribution model for site impact studies, Journal of Transportation Engineering, ASCE 131 (4) (2005).

[24] W. Wang, L. Demsetz, Application example for evaluating networks considering correlation, Journal of Construction and Management 126 (6) (2000) 467–474.

[25] H. Melhem, Y. Cheng, D. Kossler, D. Scherschligt, Wrapper methods for inductive learning: example application to bridge decks, Journal of Computing in Civil Engineering, ASCE 17 (1) (2003) 46–57.

[26] D. Bertsekas, J. Tsitsiklis, Introduction to probability, Athena Scientific, June 2002, pp. 121–154.

[27] J. Russell, Underwriting process for construction contract bonds, Journal of Management in Engineering, ASCE 8 (1) (1992) 63–80.

[28] G. John, Enhancements to the data mining process, PhD thesis, Computer Science Department, School of Engineering, Stanford University, 1997.

[29] K. Cox, S. Erick, G. Wills, R. Brachman, Visual data mining: recognizing telephone calling fraud, The Third Conference on Data Mining and Knowledge Discovery, Springer, 1997, pp. 185–192.

[30] F. Ginannotti, G. Manco, D. Pedreshi, F. Turini, Experiences with a logic-based knowledge discovery support environment, Proceedings of SIGMOD'99 Workshop on Research Issues on Data Mining and Knowledge Discovery, Philadelphia, PA, 1999.

[31] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993, pp. 46–50.

[32] S. Weiss, C. Kulikowski, How to Estimate the True Performance of a Learning System, Morgan Kaufmann Publishers, Inc., San Mateo, California, 1991, pp. 41–78.

[33] S. Haykin, T. Bhattacharya, Adaptive radar detection using supervised learning networks, Computational Neuroscience symposium, Indiana University-Purdue University at Indianapolis, 1992, pp. 35–51.

[34] D.H. Wolpert, W.G. Macready, No free lunch theorems for search, Santa Fe Institute, Technical report, No. SFI-TR-95-02-010.