Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

# Knowledge discovery and supervised machine learning in a construction project database

H. Kim, & L. Soibelman
*Department of Civil and Environmental Engineering,*
*University of Illinois at Urbana-Champaign, USA.*

## Abstract

The construction industry is experiencing explosive growth in its capability to generate and collect data. Advances in data storage technology have allowed the transformation of an enormous amount of data into computerized database systems. Nowadays, there are many efforts to convert the large amounts of data into useful patterns or trends. Knowledge Discovery in Database (KDD) is a process that combines Data Mining (DM) techniques from machine learning, pattern recognition, statistics, databases, and visualization to automatically extract concepts, interrelationships, and patterns of interest from a large database. By applying KDD and DM to the analysis of construction project data, this paper presents the results of a research that discovers the knowledge through KDD process to better identify recurring construction problems.

## 1 Introduction

Nowadays the explosive growth of many business, government, and scientific databases has far outpaced our ability to interpret and digest the available data. Such volumes of data clearly overwhelm traditional methods of data analysis such as spreadsheets and ad-hoc queries. Traditional methods can create informative reports from data, but cannot analyze the contents of those reports. Thus, a significant need exists for a new generation of techniques and tools with the ability to automatically assist humans in analyzing the mountains of data for useful knowledge (Soibelman & Kim, 2002).

As the construction industry is adapting to new computer technologies in terms of hardware and software, computerized construction data are becoming more and more available. However, in most cases, these data may not be used, or even properly stored. Several reasons exist: (i) construction managers do not have sufficient time to analyze the computerized data, (ii) complexity of the data analysis process is sometimes beyond the simple application, and (iii) up to now, there is no well defined automated mechanism to extract, preprocess and analyze the data and summarize the results so that the site managers can use it. In this paper, specific issues to consider during the KDD process on construction databases are presented since the complexity of construction data (limited breadth or coverage, data outliers, diverse forms of data, high dimensionality, etc.) makes development of an appropriate KDD difficult.

## 2 KDD and data mining

Historically, the notion of finding useful patterns in raw data has been given various names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing (Fayyad et al. 1996). KDD can be considered an inter-disciplinary field involving concepts from machine learning, database technology, statistics, mathematics, high performance computing and visualization.

### 2.1 KDD framework

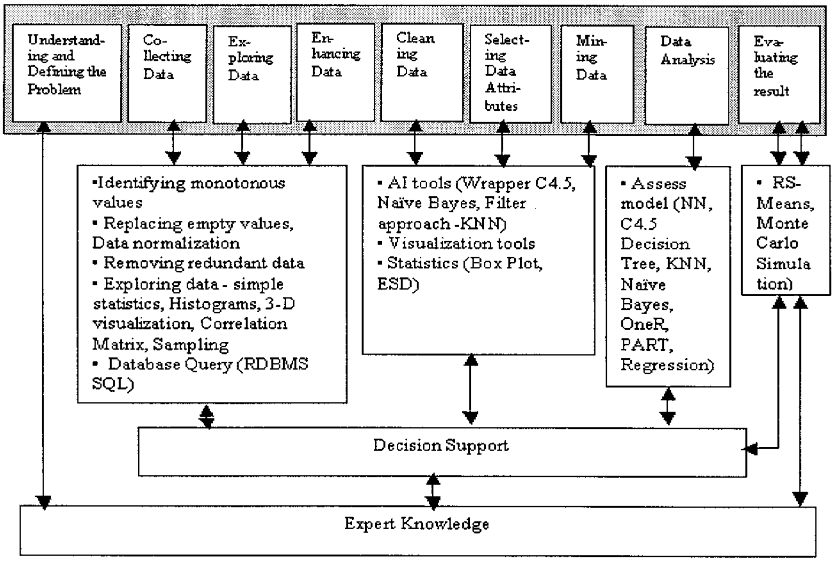
Understand-ing and Defining the Problem
Co-llecting Data
Ex-ploring Data
En-hancing Data
Clean ing Data
Select-ing Data Attri-butes
Min-ing Data
Data Analysis
Eva-luating the result
•Identifying monotonous values
• Replacing empty values, Data normalization
• Removing redundant data
• Exploring data - simple statistics, Histograms, 3-D visualization, Correlation Matrix, Sampling
• Database Query (RDBMS SQL)
• AI tools (Wrapper C4.5, Naïve Bayes, Filter approach -KNN)
• Visualization tools
• Statistics (Box Plot, ESD)
• Assess model (NN, C4.5 Decision Tree, KNN, Naïve Bayes, OneR, PART, Regression)
• RS-Means, Monte Carlo Simula tion)
Decision Support
Expert Knowledge

Figure 1. Entire KDD process (Kim, 2002)

Based on guidelines and strategies obtained from literature survey, a KDD framework was developed on five case studies. The KDD framework presented in Figure 1 differs from existing KDD process in that it puts more emphasis on data preparation process due to the unique characteristics in construction databases such as manual data entering, many different factors in a project, and no standard for data exchange, etc. The KDD framework built in this research combines technologies such as statistics, machine learning, database technology, and data visualization to discover construction knowledge from construction data.

## 3 Case studies

Five case studies were conducted to complete/validate the proposed KDD process framework by checking the feasibility in the construction database as shown in Table 1. This paper describes one construction project to illustrate the detailed steps in each process. However, overall results from each KDD application in five case studies were compared to find the most accurate data mining tools. All case studies proved to be successful in terms of generating construction knowledge in different kinds of construction databases.

Table 1. List of case studies applied to knowledge discovery framework

| Ongoing construction project | Finished construction projects | |
|---|---|---|
| | Projects with large numbers of data | Projects with small numbers of data |
| - Flood Control Project | - Hospital rehabilitation project<br>- Ammunition supply facility project | - Perimeter fence security project<br>- Material process facility project |

**Step 1: Identifying problems**

The first step of KDD implementation is the identification of problems. Therefore, the domain information of construction delays had been obtained before any data was analyzed. The initial data survey for a project in Fort Wayne, Indiana, provided by US Corps of Engineers, demonstrated that one activity called "Installation of drainage pipelines (sub-activities: excavating the ground, installing pipelines, backfilling compacted, and Erosion protection)" was behind schedule in 54% of the instances (120 out of 224).

**Step 2: Data collecting**

Data stored in a database such as Oracle, SQL server or MS ACCESS needs to be converted into a text file through SQL statements so that KDD algorithms can understand them. SQL contains facilities for defining, manipulating, and controlling relational databases. In this research SQL statements were used to create a table that relates various inputs such as activity_id, feature_description ("excavating the ground", "installing pipeline", "backfilling compact material", and "erosion protection"), problem ("rework", "inaccurate drawing", "shortage of equipment" and so on), and delay_prediction (delay_prediction = "yes" or "no"). Figure 2 shows the architecture of collecting relevant data through SQL statements.

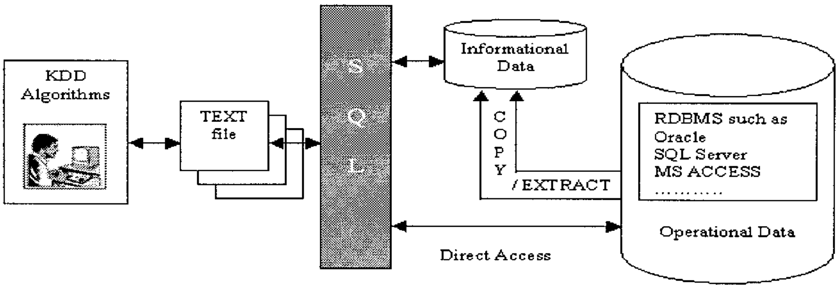

Figure 2. Overall procedures of data extraction from SQL queries

**Step 3: Data exploring**

In some cases, the optimal strategy is to work with an entire database, but this can be cost-prohibitive. Therefore, it is important to know what is stored in the database. Construction databases usually have scores of tables that are linked together in relational databases with relevant information scattered among them, rather than having a single database table. In the case study database, for example, each activity had its own ACTIVITY DESCRIPTION, AMOUNT, LOCATION, STARTING/FINISHING DATE, and DAILY COMMENT. In identifying the causes of construction delays, descriptions from the tables such as DAILY COMMENTS, QUALITY ASSURANCE, QUALITY CONTROL, and DAILY REVIEW on each activity seemed to be important, because often activities with similar comments tend to produce similar patterns. It is also important to consider that third-party databases can help to add more data to the existing database (weather information such as rainfall and temperature and construction material price index).

After exploring the database, two common problems have been noted: special data values with special semantics, and unmatched time. One interesting error found in the database was that certain activities had the value of '999' in total float. After investigating on the data, real values of these records were meant to be unknown rather than a high number of total float and removed from the database. Another common error that is found in the database was the type of time mismatch. The entire database was examined and one interesting pattern was found: for a certain period, virtually none of the activities had been delayed. However, the database had been given to us before the activities took place, thus those data had no opportunity to be filled out. After investigating the activities, it turned out that the activities had been recorded before those activities took place by accident. Those data were removed since they did not hold valid information.

**Step 4: Data enhancing**

The process of data enhancing is one of the most important steps of the entire process, and one of the most time-consuming and difficult. Key benefit of this step is that the data enhancing process prepares both the data and the analyst.

When data is properly prepared, the analyst unavoidably gains understanding and insight into the content, range of applicability, and limits to use of the data.

In the case study database, there were several items that needed to be enhanced before data analysis. For example, weather data were left blank for each workday in RMS. Since weather is closely related to construction activities, adding more weather data to the system helped to extract important patterns. Also, the database was composed of several reports and comments that needed to be categorized into a certain number of features. Many instances of missing and empty values were found. Therefore, proper ways of enhancing data need to be implemented (Kim, 2002). In this research, missing values were replaced with most probable values and categorical data were converted to numerical data since not all machine learning tools can understand categorical data. Also, monotonous values were identified/deleted since it does not provide any useful information.

**Step 5: Data cleaning**

The results (Table 2) from the case study demonstrated that in statistical methods of automatic detection, boxplot achieved a better average error rate (19.6%) to make a prediction on construction delays for future project, compared to the result of ESD (25.7%). However, the average error rate of detection from domain knowledge was 17.8% which means that by eliminating data as outliers arbitrarily, useful information was also removed from the population. When the authors investigated on those data, it was found that those 14 data out of 248 instances (5.64%) were stored incorrectly into the database, judging from the fact that some of the activities in the database were finished at the same time when the project was finished. It turned out that the database administrator of the project had not entered actual finishing dates into the database for some activities accidentally when the actual construction was completed. However, when the whole construction project was finished, the system was built in a way that automatically finishes all the activities at the finishing date of the construction project, even when some of the activities were not recorded to be completed. This would be one of the reasons why it took about 150 days for a certain construction activities when it was supposed to be finished in 1 to 4 days. Outliers identified were removed since those data did not add any information to the whole dataset.

Table 2. Comparison of outlier identifications

| | Method name | No. of outliers identified | NN error rate (%) |
|---|---|---|---|
| **Original Dataset** | - | 0 | 31.5 |
| **Automatic Methods** | Boxplot | 26 | 19.6 |
| | ESD | 10 | 25.7 |
| **Non-Automatic Method** | Domain knowledge | 14 | 17.8 |

**Step 6: Selecting data attributes**

Construction databases consist of large amounts of data that includes labor, cost material, schedule, resources, etc. Therefore, if we include just one or two items to predict any estimated values, the results might not be correct since not all the factors were considered. However, once we decide to include all the factors, it gets difficult to know where to put focuses on.

This research attempted to apply feature selection methods to identify important factors for construction dalays. There are many possible measures for evaluating feature selection algorithms. Authors used the following criteria to compare the performance of the feature subset selection algorithm:

1) Dimensionality reduction: The amount of dimensionality reduction of each method was counted. The number of reduction is considered to be important since the large number of potential features might constitute a serious obstacle to efficiency of most learning algorithms

2) Classification accuracy: The error rate of all models has been measured. By removing irrelevant or noisy attributes, feature selection methods are assumed to decrease the error rates of learning algorithms. Accuracies are measured to find the best algorithm between filter and wrapper approaches in this section.

3) Computing time of feature selection: The time of feature selection was measured. The less amount of computing time is preferred due to its simplicity.

**Comparison of filter and wrapper approaches**

- Filter and wrapper approaches have been tested. Wrapper approach considers as relevant only 21 % of features while filter removes a smaller number features (average dimensionality reduction of 62%).
- Wrapper approach appears to be a better method of feature selection than filter approach since the accuracy rate (95.4±2.3%) of C4.5 in wrapper approach was much higher than the accuracy rate in filter approach (89±2.3%). From Naïve Bayes, the accuracy of filter approach (78±5.8%) was improved to the accuracy (80±2.2%) of wrapper approach.
- Even though wrapper approach gives better predictive accuracy, wrapper approach is limited by the time constraint. As the number of testing features gets larger, the number of computation increases exponentially ($2^n$, where n denotes the number of attributes to be tested).

Testing of feature selection methods in this research produced almost the same results when ignoring all the attributes of the significant level less than 2% while considering the seven remaining attributes of filter approach to be redundant. Thus, nine attributes were chosen to be important from two different approaches (filter/wrapper).

Figure 3 shows the comparisons of average error rates, computed in five case studies as they are improved in data preparation that includes data enhancing, data cleaning, and data selection. With the proper data preparation, Figure 3 shows that there is a significant accuracy implementation by proper data preparation (data enhancing, data cleaning, and data selection).

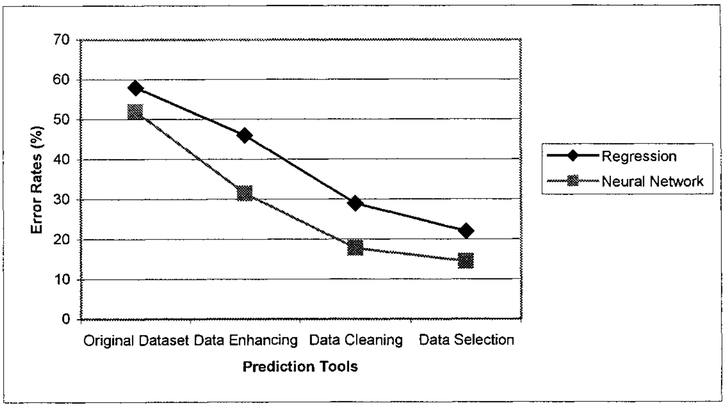

Figure 3. Increase of error rates through data preparation process

## Step 7 & 8: Data algorithm/analysis

This is the step in KDD process that requires particular data mining algorithms to produce a particular enumeration of patterns over the database under some acceptable computational efficiency limitations. With the development and penetration of data mining within different fields, many data mining algorithms emerged. This research utilized several different classification algorithms (C4.5 decision tree, Naïve Bayes, PART, and One-R) and prediction algorithms (regression, Neural Networks, and k Nearest Neighbor) to generate construction knowledge and results were compared to identify the best algorithm in this research.

## C4.5 Decision tree

In Figure 4, each node contains information about the number of instances and percentages of that node, and about the distribution of dependent variable values. The instances at the root node are all of the instances in the training set. This node contains 224 instances, of which 54 percent are instances of delay and 46 percent are of no delay. Below the root node (parent) is the first split that, in this case, splits the data into two new nodes (children), based on whether "Inaccurate Site Survey" is yes or no. C4.5 decision trees produced is shown in Figure 4. A tree that has only pure leaf nodes is called a pure tree. Most trees are impure, that is, their leaf nodes contain cases with more than one outcome. Figure 4 reveals the following interesting patterns:

- Weather, considered responsible for delays by site managers, appears not to be the most important cause in determining delays.
- Activities with Inaccurate Site Surveys are always delayed in the schedule.

Also, Shortage of Equipment, Seasons, and Incomplete Drawing are very significant factors in determining schedule delays since inductive algorithms tried to prioritize their splits by choosing the most significant split first.

Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

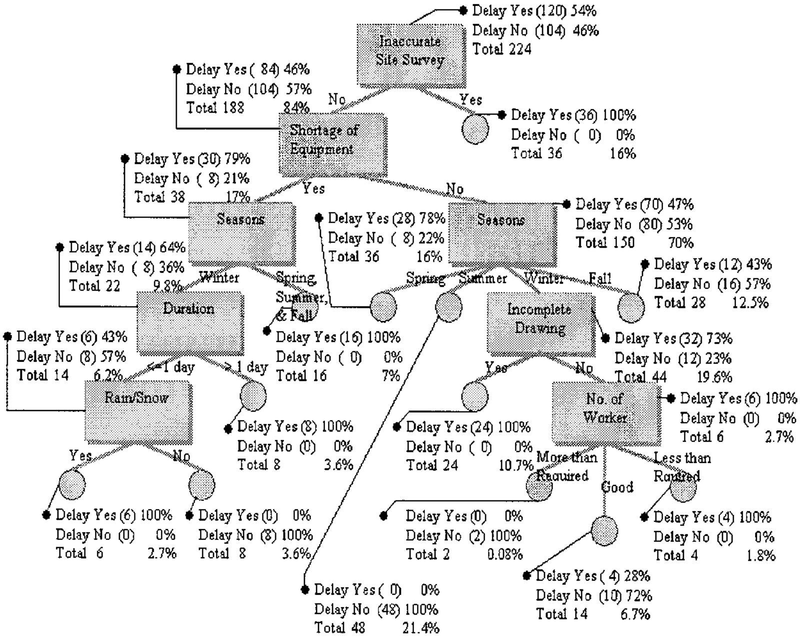

Figure 4. Decision tree of schedule delays on drainage pipeline activities

## Overall results

This section attempts to determine which of a number of investigated classification/prediction techniques provides the best classification, based on a limited data set of RMS construction data. Supervised classification techniques including One-R, C4.5 decision trees, and PART can be used to classify the construction activity delay.

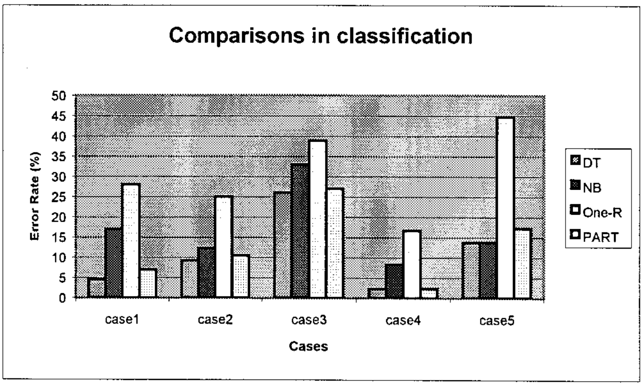

Figure 5. Overall performance of classification tools

Their accuracies have been compared on the basis of the error rates achieved. Error rates as low as 4.78 % were achieved as shown in Figure 5.

Prediction techniques including multiple regression, k nearest neighbor, and neural networks have been used to predict the construction delay. Error rates as low as 14.713 % were achieved as shown in Figure 6. The standard deviations ranged from 4.5 % to 8.7 %. Also, it is found that the best accuracy of classification was achieved from C4.5 decision tree and that the best accuracy of prediction was obtained from neural network.

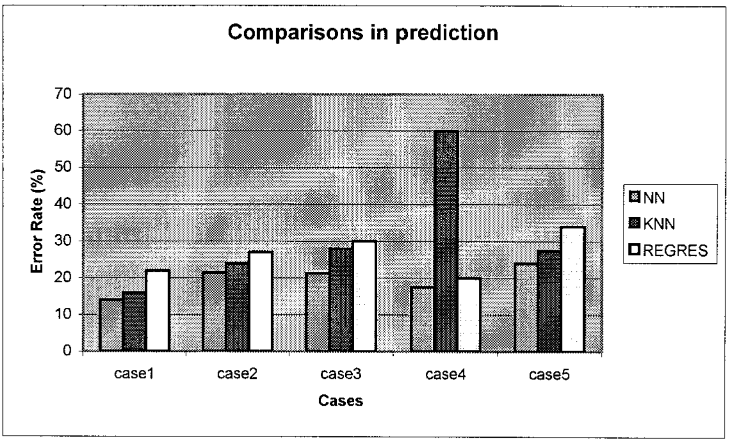

Figure 6. Overall performance of prediction tools

**Step 9: Preliminary evaluation**

According to the preliminary results of this case study, the main cause of activity delays in the Flooding Control Project at Fort Wayne was "Inaccurate site survey" rather than the weather related problems initially assumed by site managers. Discussions with site managers in the project confirmed the importance of equipment, such as ground penetration radar, to make the site surveys more accurate. Ground Penetrating Radar (GPR) is an equipment to locate existing underground pipelines and construction structures. It is considered to be very cost-effective in the flood control project, since one of the most frequent problems in activity delays was due to "Inaccurate site survey." The potential savings to be obtained after buying a GPR was calculated for Fort Wayne project. We were able to predict a significant amount of savings if compared to $15,000 investment to buy the GPR equipment. The $587,391 savings figure was obtained by the multiplying the daily construction cost by the expected number of instances (75) for the activity of drainage pipeline installation during the next stage of the construction process by the number of days to be saved through using the GPR (0.83 days) and by the previous probability of construction delay (0.16) [$93,982.56= 9,436 * 75 * 0.83* 0.16]. The number of days to be saved from the GPR was obtained by running the NN with the weights learned from the previous phase of the same project. In addition, the expected savings would increase even more if schedule related penalties for delays were considered.

## 4 Conclusions

With the use of large database, this research utilizes KDD technologies that reveal predictable patterns in construction data that was previously thought to be chaotic. This research applied KDD process to analyze the data to extract knowledge or patterns from the database so that the project manager may have a better understanding of causal relationships in a project. In this paper, the authors discussed an approach for discovering some useful knowledge from large amounts of data that were generated during a construction project. The proposed approach helped to guide the analysis through the application of diverse discovery techniques. Such a methodological procedure helped us to address the complexity of the domain considered and therefore to optimize our chance to discover valuable knowledge.

During the knowledge discovery approach, one of the most important, time-consuming and difficult parts of KDD process was data preparation. Domain knowledge and good understanding of the data is key to successful data preparation.

In this research, five case studies were conducted by the authors to identify the causes of construction delays. But its possible applications can be extended to different areas such as identifying the causes of cost overrun, or quality control/assurance from the RMS database. The research of KDD process to large construction data is continuously being refined and more case studies are to be followed. Eventually, knowledge discovery model-building framework developed by this research will be used to guide novice construction model builders through the process of creating models based on their own construction data.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under grant number 0093841. The authors would like to thank Mr. Francois Grobler and the US ARMY Corps of Engineers Research Laboratory for providing us access their RMS database.

## References

[1] Kim, H., Knowledge discovery and machine learning in construction project databases, PhD thesis, University of Illinois at Urbana-Champaign, 2002.

[2] Fayyad, et al. From Data Mining to Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, pp. 1-34, 1996

[3] Soibelman, L. & Kim, H., Data preparation process for construction knowledge generation through Knowledge Discovery in Databases, *Journal of Computing in Civil Engineering*, pp. 39-48, 2002.