Data Mining tools and techniques in construction by Knowledge Areas: State of the Art situation

Danilo Arba

ABSTRACT

Managing Project Controls, both from an owners perspective as from a contractor organization, is commonly, and as per the Compendium of the Guild of Project Controls, divided in 12 modules that make excellent use of Information technologies, generating an incredible amount of data, making analyzing and organizing the data crucial for the effective use of the data to improve initiating, planning, controlling and closing in the construction sector.

This paper aims to establish which data mining tools and techniques are best for use in construction to help to maximize opportunities and reduce risks in the 12 modules of the Compendium of the Guild of Project Controls.

Project Managers and Project Controls professionals can leverage the use of the validated data to ensure better decision making, faster and with the most significant benefit to the projects.

KEYWORDS: Data Mining, Knowledge Area, KDD process, Classification, Big Data, Business Intelligence, Enterprise Data Mining, Correlation, Knowledge Area, Data Validation

INTRODUCTION

"Hope is NOT a valid management Strategy". History is studded with denial and project failures around the world, with "a significant number of the construction project have encountered problems during the construction phase, 98% of projects incur in cost overruns or delays", related to the fact that "risk has not been dealt with adequately, that subsequently resulted in a limited performance in the built of the project with increasing cost and time delays."

There are various theories and methods to manage data in the construction industry, and with the help of new technologies, the construction industry is adapting and start to implement it,

---

5 Bizon-Górecka J, Górecki J. Risk of the construction investment project in perspective of the execution model. Studies & Proceedings of Polish Associations for Knowledge Management. 2015;74:4-15
such as using of data mining to manage the various knowledge areas more productively and efficiently. The use of data mining can discover patterns on construction risks from raw data in a previously unknown way.

Two items are essential in this research, data mining tools and techniques and how this relates to the knowledge areas to achieve project success based on the specifications we have set-up for our project or product or both.

In a general way, risk/opportunity can be "defined as decisions to accept exposure or reduce vulnerabilities by either mitigating the risk or applying adequate controls"\(^6\), And this should be associated primarily with the measure of deviation from the pre-planned values, and it is usually defined as the probability of those deviations.

In more detail risk can be defined as "A probability or threat of damage, injury, liability, loss, or any other negative occurrence that is caused by external or internal vulnerabilities, and that may be avoided through preemptive action."

How can we define data mining? It is like a type of deep technology analysis of data, which can establish a forecasting model instead of a back retrospective model. Traditionally statistical analysis tools are used to test the past situation while data mining technology aimed at finding unexpected relationships through discoverable, predictable, pattern matching algorithms, traditionally statistical analysis tools used to test the previous case.

On the other hand, risk management is of significant importance for achieving the project objectives, not only minimizing bad results but also acting as a guide to maximizing the positive results. If we analyze the current state of knowledge on risk management, we can state that the necessity of considering risk in project management has been well established [4], although not correctly implemented in the construction industry and management tools have not been appropriately applied. Heuristic models are still of dominant importance, in risk management and attempts at modelling refer to the evaluation of the behavior of some parameters exposed to random effects.

As of today, we are looking inside the natural phenomena that affect the risk levels and diagnosing the magnitude of the risk-based on empirical research is an open research field.

To achieve the objective of a given project, risk analysis can provide the foundations for building the confidence of accomplishing them. A more comprehensive approach is needed and thinks not only about the weak links of primary nature, which occur in its respective processes, but also about a consequent risk which is the effect of mutual relations of those processes, that will

---


\(^7\) When was the last time you said this? (n.d.). BusinessDictionary.com. https://www.businessdictionary.com/definition/risk.html
develop more dynamically. One must also remember to factor in risk management the treats that/can occur in the future. Project Directors and Project Managers face continually new and changing challenges that contribute to changes in project management.8

In the particular field of construction, the amount of data available, structured or unstructured, is immense and rarely properly analyzed, utilized or archived to make sound management decisions and manage risks. Data mining technology can be used to manage risks in a new and more efficient way, while technology allows us to maintain an ever-increasing amount of data, with an increase of 40% year on year while form from 2011 to 2016 it increased nine-fold.9 We will focus in this paper to explain how, through data mining, we can increase the chances of success in a construction project, with a real retail construction project. We will apply data mining to find useful relationships between risk factors, remedial measures and results, to assess the risk combination, and to help the decision-makers to find effective means to avoid the loss of construction risk.

The focus of this technical paper is which data mining techniques are more appropriate in risk management and how this can be related to increased chances of success in the construction of major construction projects related to the project and not the end product.

What are we trying to find out? We will review 16 of the standard data mining tools and techniques to find out which:

1. Best data mining tool/technique compared to the 12 modules of the Compendium of the Guild of Project Control.
2. Compared to the 4 typical levels of Artificial Intelligence which data mining tools have the most significant positive impact for each module of the Compendium of the Guild of Project Controls
3. Which data mining tools have the most significant benefit in managing risk and exploiting opportunities?

METHODOLOGY

Step 1

The development of data mining has been powered by the desire to find meaningful, new patterns in a real-world database. It is commonly applied in many fields such as shopping bag analysis, insurance, fraud detection, and text categorization. Unfortunately, there were few issues to argue about the application of data mining in construction risk management. So, it is worthwhile to experiment in this field to manage the risk in the construction project.

Fig.02: KDD Process in Data Mining

---

The knowledge discovery process, in fig.02, is interactive and iterative and includes nine steps. At each stage, the process is iterative, that could mean moving back to the previous actions might be necessary. This is necessary to understand the solutions in each node. It is starting by determining the objectives of KDD and ending implementing the knowledge discovered. Data mining can now begin, and after this, we can make changes in the application domain.

We have identified 16 possible data mining tools & techniques\(^{12}\) that we can then compare to the 12 modules of the GCP, as per below picture:

---


Step 2

We can define 16 data tools and techniques and definitions\textsuperscript{14} as feasible alternatives;

1. **Data Cleaning & preparation**, Data cleaning and preparation is a vital part of the data mining process. Raw data must be cleansed and formatted to be useful in different analytic methods. Data cleaning and preparation includes various elements of data modelling, transformation, data migration, ETL, ELT, data integration, and aggregation. It is a necessary step for understanding the basic features and attributes of data to determine its best use. The business value of data cleaning and preparation is self-evident. Without this first step, data is either meaningless to an organization or unreliable due to its quality. Companies must be able to trust their data, the results of its analytics, and the action created from those results. These steps are also necessary for data quality and proper data governance.

2. **Tracking Patterns**, Tracking patterns is a fundamental data mining technique. It involves identifying and monitoring trends or patterns in data to make intelligent inferences about business outcomes. Once an organization identifies a trend in sales data, for example, there is a basis for taking action to capitalize on that insight. If it is determined that a specific product is selling more than others for a particular demographic, an organization can use this knowledge to create similar products or services, or only better stock the original product for this demographic.

3. **Classification**, this analysis used to retrieve valuable and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

4. **Association Rules**: This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. **Outer Detection**, This type of data mining technique refers to the observation of data items in the dataset which do not match an expected pattern or expected behaviour. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection. Outer detection is also called Outlier Analysis or Outlier mining.

6. **Clustering**, Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

7. **Regression**, Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to determine the likelihood of a specific variable, given the presence of other variables.

8. **Prediction**, Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification. It analyzes past events or instances in the right sequence for predicting future development.

9. **Sequential Patterns**, this data mining technique helps to discover or identify similar patterns or trends in transaction data for a specified period.

\textsuperscript{14} 16 data mining techniques: The complete list. (2019, July 29). Talend Real-Time Open Source Data Integration Software. \url{https://www.talend.com/resources/data-mining-techniques/}
10. **Decision Trees**, Decision trees are a specific type of predictive model that lets organizations effectively mine data. Technically, a decision tree is part of machine learning, but it is more popularly known as a white box machine learning technique because of its extremely straightforward nature. A decision tree enables users to understand how the data inputs affect the outputs clearly. When various decision tree models are combined, they create predictive analytics models known as a random forest. Complicated random forest models considered black-box machine learning techniques because it is not always easy to understand their outputs based on their inputs. In most cases, however, this basic form of ensemble modelling is more accurate than using decision trees on their own.

11. **Statistical techniques**, Statistical techniques are at the core of most analytics involved in the data mining process. The different analytics models based on statistical concepts, which output numerical values that apply to specific business objectives. For instance, neural networks use complex statistics based on different weights and measures to determine if a picture is a dog or a cat in image recognition systems. Statistical models represent one of two main branches of artificial intelligence. The models for some statistical techniques are static, while others involving machine learning get better with time.

12. **Visualization**, Data visualizations are another essential element of data mining. They grant users insight into data based on sensory perceptions that people can see. Today's data visualizations are dynamic, useful for streaming data in real-time, and characterized by different colours that reveal different trends and patterns in data. Dashboards are a powerful way to use data visualizations to uncover data mining insights. Organizations can base dashboards on various metrics and use visualizations to visually highlight patterns in data, instead of merely using numerical outputs of statistical models.

13. **Neural Networks**, A neural network is a specific type of machine learning model that is often used with AI and deep learning. Named after the fact that they have different layers which resemble the way neurons work in the human brain, neural networks are one of the more accurate machine learning models used today. Although a neural network can be a powerful tool in data mining, organizations should take caution when using it: some of these neural network models are incredibly complex, which makes it difficult to understand how a neural network determined an output.

14. **Data Warehousing**, Data warehousing is an essential part of the data mining process. Traditionally, data warehousing involved storing structured data in relational database management systems so it could be analyzed for business intelligence, reporting, and necessary dashboarding capabilities. Today, there are cloud data warehouses and data warehouses in semi-structured and unstructured data stores like Hadoop. While data warehouses are traditionally used for historical data, many modern approaches can provide an in-depth, real-time analysis of data.

15. **Long-term Memory Processing**, Long term memory processing refers to the ability to analyze data over extended periods. The historical data stored in data warehouses is
useful for this purpose. When an organization can perform analytics on an extended period, it is able to identify patterns that otherwise might be too subtle to detect. For example, by analyzing attrition over several years, an organization may find subtle clues that could lead to reducing churn in finance.

16. **Machine learning and artificial intelligence**, Machine learning and artificial intelligence (AI) represent some of the most advanced developments in data mining. Advanced forms of machine learning like deep learning, offer highly accurate predictions when working with data at scale. Consequently, they are useful for processing data in AI deployments like computer vision, speech recognition, or sophisticated text analytics using Natural Language Processing. These data mining techniques are suitable for determining value from semi-structured and unstructured data.

**Step 3**

Based on our 16 data mining tools and techniques and 12 modules from the Guild of Project Controls Compendiums, we can compare how each data mining tools/techniques compares to the relevant module of the Compendium.

---

**Fig.04:** Comparison of data mining tools vs GCP\(^\text{15}\) using the levels of Artificial Intelligence\(^\text{16}\)

---

\(^\text{15}\) Comparison chart data mining tools and techniques vs Guild of Project Controls Compendium, by the author  
\(^\text{16}\) Reynoso, R. (n.d.). Four main types of artificial intelligence. Learning Hub | G2. [https://learn.g2.com/types-of-artificial-intelligence](https://learn.g2.com/types-of-artificial-intelligence)
We have considered the 4 typical levels of Artificial Intelligence\(^{17}\). We can use four types of artificial intelligence that comprise smaller aspects of the general realm of AI.

1. Type 1: Reactive Machines (Score 1 Point in Figures 4 and 5)
2. Type 2: Limited Memory (Score 2 Points in Figures 4 and 5)
3. Type 3: Theory of Mind (Score 3 Points in Figures 4 and 5)
4. Type 4: Self-Awareness (Score 4 Points in Figures 4 and 5)

While we can consider AI as undoubtedly multifaceted, there are specific types of artificial intelligence under which extended categories fall. Even if we can find a variety of terms and definitions in AI that can make it challenging to navigate the difference between classes, subsets, or types of artificial intelligence – and no, they are not all the same.

Some subsets of AI include machine learning, big data, and natural language processing (NLP); however, in this technical paper, we are covering the four main types of artificial intelligence: reactive machines, limited memory, a theory of mind, and self-awareness, which are also our scoring points in figure 4 and 5.

To have a better clear view, we created a heat map by row (tools and techniques) and by column (GPC modules). Firstly we will rank order it by the GPC modules;

\(^{17}\) Reynoso, R. (n.d.). Four main types of artificial intelligence. Learning Hub | G2. [https://learn.g2.com/types-of-artificial-intelligence](https://learn.g2.com/types-of-artificial-intelligence)
From figure 5, we can analyze the following modules are the greatest impacted:

1. Estimating and Budgeting
2. Planning and Scheduling
3. Risk and Opportunity

As soon as we execute this, we can rank order the data by the row, but leaving the modules in the first rank, the data mining tools and technique, from the highest to the lowest but leaving the modules ranked from 1 to 12.

18 Heat map of the tools and technique and GPC modules, by author
We can easily see which of the following data mining tools and technique have the most significant impact overall on the modules;

1. Long-term memory processing
2. Data cleaning and preparation
3. Data warehousing

---

19 Heat map of the tools and technique ordered from highest to lowest, by author
Step 4

Now we have all the elements to analyze which data mining tool and technique and which module from the GCP compendium is impacted and by which type of artificial intelligence type has the greatest and least impact, this will allow us to properly understand which are the answers to our technical papers questions.

FINDINGS

Step 5

The scoring points are the result of the author experience and evaluation, and we can summarize them as for below figure using Multi-Dimensional Decision Making (MDDM) using dominance

<table>
<thead>
<tr>
<th>DATA MINING TOOLS &amp; TECHNIQUES</th>
<th>Mod 8</th>
<th>Mod 7</th>
<th>Mod 4</th>
<th>Mod 1</th>
<th>Mod 12</th>
<th>Mod 9</th>
<th>Mod 11</th>
<th>Mod 6</th>
<th>Mod 10</th>
<th>Mod 2</th>
<th>Mod 3</th>
<th>Mod 5</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Long-term memory processing</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1. Data cleaning and preparation</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>14. Data warehousing</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>9. Sequential patterns</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2. Tracking patterns</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>12. Visualization</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>16. Machine learning and artificial intelligence</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>4. Association</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>7. Regression</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>8. Prediction</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>3. Classification</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>11. Statistical techniques</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>5. Outer detection</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>10. Decision trees</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>6. Clustering</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>13. Neural networks</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>

Fig.07: Multi-dimensional decision-making results using dominance

Form our figure 7, it is clear which module is most advanced and which tool and technique are mostly used, so we can determine that:


21 Multi-dimensional decision-making results using dominance, by author
The most advanced modules are Estimating and budgeting, Planning & Scheduling and Risk & Opportunity.

The most used tool and techniques are Long-Term memory processing, data cleaning and preparation, data warehousing.

Based on this, we can now answer our previously discussed problems:

1. Best data mining tool/technique compared to the 12 modules of the Compendium of the Guild of Project Control? As per figure 7 above, we can see that the three best data mining tool/technique compared to our 12 modules of the Compendium of GPC are:
   i. Long-term memory processing
   ii. Data Cleaning and Preparation
   iii. Data warehousing

2. Compared to the 4 typical levels of Artificial Intelligence which data mining tools have the most significant positive impact for each module of the Compendium of the Guild of Project Controls?
   a. As we can see from the fig.07 above we can see that the first modules are mostly impacted by the theory of mind and self-awareness, and as we go to the least impacted modules we can see that the most valuable artificial intelligence types are reactive machines and limited memory which are the two most basic types of artificial intelligence.

3. Which data mining tools have the most significant benefit in managing risk and exploiting opportunities?
   a. As we can see from the above figure, 7 tracking patterns and regression is the data mining tool with the most significant benefit to managing risk and opportunities, while the theory of mind and self-awareness are the two most essential levels of artificial intelligence used.

CONCLUSIONS

Independently if we are the owner or the contractor, we can control the results of our analysis by using Pareto Charts. The Pareto Principle, or the 80/20 rule, states that for many phenomena, 80% of the effect comes from 20% of the effort. The principle has been named after Vilfredo Pareto—an Italian economist—who, back in 1895, noticed that about 80% of Italy's land belonged to 20% of the country's population.

---

When we apply this principle to the context of risk/opportunity monitoring and control, and to the above results, we can evaluate the as-is (before improvement) situation that we will be able to compare with a following application of the same principle. The author suggests reviewing and compare the results after one or two years and analyze if any changes occur.

Then a "before" and "after" analysis needs to be performed to verify if the interventions or corrective actions solved the problem, or did it make things worse?

**Fig.08: Result of the Pareto analysis on the Guild of Project Controls Compendium**

---

25 Result of the Pareto analysis on the Guild of Project Controls Compendium, by author
Fig.09: Result of the Pareto analysis on the 16 data mining tools

26 Result of the Pareto analysis on the 16 data mining tools, by author
BIBLIOGRAPHY


[14] Comparison chart data mining tools and techniques vs Guild of Project Controls Compendium, by the author


[16] Heat map of the tools and technique and GPC modules, by author

[17] Heat map of the tools and technique ordered from highest to lowest, by author


[19] Multi-dimensional decision-making results using dominance, by author


[23] Result of the Pareto analysis on the Guild of Project Controls Compendium, by author

[24] Result of the Pareto analysis on the 16 data mining tools, by author
About the Author

Danilo Arba

Milan, Italy
Lima, Peru

Danilo Arba is a project controls & management enthusiast, with 20 years of experience. He is a Certified Cost Engineer with an Executive MBA from Politecnico di Milano. With a thorough understanding of EPC (Engineering, Procurement, and Construction) industry, he has a verifiable track record of planning multimillion/billion-dollar construction projects worldwide. He lived & worked all his life around the world from South America, Africa, South East Asia to Europe. He is adept at building and leading cross-functional teams from project conception to completion, optimising performance, contractual, and financial deliverables. Currently he is furthering his education by way of a distance learning mentoring course, under the tutorage of Dr Paul D. Giammalvo, CDT, CCE, MScPM, MRICS, GPM-m Senior Technical Advisor, PT Mitrata Citragraha, to attain Guild of Project Controls certification.

Danilo lives in Milan, Italy and Lima, Peru and can be contacted at danilo.arba@mip.polimi.it