A Computational Framework for Railway Incident Analysis: from Data Mining to Data Visualization

Completed Research

Sérgio Viademonte
Instituto Tecnológico Vale
sergio.viademonte@itv.org

Cleidson de Souza
Universidade Federal do Pará
cleidson.desouza@acm.org

Nikolas Carneiro
Instituto Tecnológico Vale
nikolas.carneiro@itv.org

Jair Ferreira Junior
Instituto Tecnológico Vale
jair.junior@itv.org

Wallace Lyra
Simon Fraser University
wpintoli@sfu.ca

Abstract

This document presents a computational framework for railways incident analysis and risk assessment. The computational framework combines data mining technologies with a visual programming framework, to generate knowledge from incident data and allows users to explore such knowledge, to obtain insights about conditions that impact railways safety. This document describes the workflow on the development of this computational framework, discusses data processing activities, the evaluation of the developed computational framework, and the visual knowledge representation capability.

Keywords

Decision support systems, data mining, machine learning, data visualization, incident analysis.

Introduction

Incident analysis is an important activity in many industries, mainly in the ones that operate in conditions that involve high-risk work and operational environments, as the mining industry. Vale is a multinational mining organization, headquartered in Brazil, with operations that spread over various countries. Besides its mining operations, Vale also controls its main logistic operations, which includes railways and maritime transports. Minerals extracted from its mines, are carried through networks of railways and maritime transports to customer sites.

Railways in Brazil, specifically the ones operated by Vale, cover long distances, running through both sparsely and densely populated areas, as well as highly remote areas. It is a risky operation, not only to employees but also to the communities living in areas along the railways. Vale adopts a range of initiatives to mitigate risk of incidents, such as training on safety procedures, the use of technology such as sensors and cameras, and best practices. Vale also collects and stores incident data for report and analysis purposes. An approach for risk mitigation is the capability of identifying information about operational conditions, which lead to incidents, and apply such information to assess the likelihood of occurrence of incidents (Fukuda 2002; Heinrich 1931), i.e. to predict such events.

The research project described in this document relates to the development of a computational framework for railways incident analysis and prediction, to support safety engineers on the tasks of...
railways incident management and risk assessment. Emphasis has been given in the development of capabilities to allow users to visualize and explore incident related data.

Our objectives are to be able to identify and describe the context in which railways incident occur, at Vale's (operated) railways. To determine if there are certain conditions that lead to incidents more often than others, i.e. patterns of incidents, and provide a computational capability to compute the likelihood of an incident occurrence.

The computational framework has been designed to discover nontrivial relationships among railways incident data, and to compute a risk index for incident occurrence. Also, an important feature of the framework is to make explicit the discovered "knowledge" (relationships), in a way that users can visually analyze such knowledge and build insights about the circumstances involved in a certain incident. We have applied a combination of data mining and machine learning technologies with visualization techniques. Therefore, the data mining stages of data selection and processing, generation of descriptive and predictive models have been applied and are discussed in this document, as well as the development of a visualization capability. Special effort has been placed on such visual analytic capability, and this is discussed throughout this document.

The contributions of the project presented in this document are in the development of a combined computer application, using data mining, machine learning and data visualization technologies. Also, an important contribution is the experience in developing this computer application for railways incident analysis, the study of feasibility of such approach for railways operation, and the identification of relevant information for this specific type of analysis. This activity highlighted the need for a broad perspective for feature selections in this domain.

This document is organized as follows; next section presents a discussion and some related works on railways incident analysis, and follows on describing the development of the computational framework, including the stages of data gathering and processing. It discusses the visual analytic capability that has been developed. At the end, it presents our conclusions and directions of future research and development.

**Railways Incident Analysis**

This document applies the words "incidents" and "accidents" interchangeably, although "accidents" can be considered a class, or type, of "incidents". Furthermore, for the reason of information privacy and confidentiality, as we have access and used de facto Vale's operational data, this document has replaced some data, values and time periods. But this does not compromise the understanding and explanations of the work developed and discussed in this document.

Accident prevention has been the subject of many studies and investigations, worth to emphasize the work of Heinrich (Heinrich 1931), which proposed the idea of common cause hypotheses for accidents. In a work on railway incident analysis in the United Kingdom (UK), three methods for investigation were applied: serious incidents were investigated via formal inquires (a panel of experts discuss a particular incident and interview the staff involved), less serious incidents were analyzed through the SPAD system (signal passed at danger) and CIRAS (confidential incident reporting and analysis system), which is normally used for near miss cases (Wright and Schaab 2004).

Fukuda (Fukuda 2002) states that many industries, including railways, energy and airlines are systematically collecting and analyzing information on incidents, with the purpose of preventing them, through their incident reporting systems (IRS). The ultimate goal is to be able to predict accidents, propose countermeasures, and mitigate risk. Despite of those efforts, the study (Fukuda 2002) concludes that there are no universal methods established for incidents analysis, and each industry develops its individual solutions. It (Fukuda 2002) proposes a conceptual information system framework, to extract incident data and stores it in databases, for further analysis, and also includes a feedback mechanism. Another interesting initiative is the work developed by Evans (Evans 2011), which presents an analysis of fatal train accident, specifically collisions and derailments, rates and trends on Europe’s main line railways from 1980 to 2009. The paper uses a set of data for the European Union together with Norway and Switzerland.
The work described by Lira (Lira et al. 2014) applies a visual analytic (VA) approach for incident analysis. It compiles an incident risk index along a railway, and visually displays the index on a geographical map together with socioeconomic information about the towns and cities associated with incidents.

The task of incidents analysis and prediction is very challenging; the research described in this document aims to develop a computational framework to support the preventive analysis of incidents. As stated by Lira (Lira et al. 2014), even if we cannot accurately predict incidents, it is possible to highlight the circumstances in which they are more prone to happen, and quantify this.

A Computational Framework For Incident Analysis And Prediction

The computational framework for incident analysis and prediction described in this document has been grounded in our previous research projects. In previous research, a computational framework applying data mining and artificial neural networks, has been designed and applied to the domains of credit scoring (Pree et al. 1997) and aviation short term weather forecast (Viademonte and Burstein 2006). This research has also been grounded on the visual analytic capability for railways incident analysis developed by Lira (Lira et al. 2014). Please, refer to specific bibliographic references for further discussions and details about those projects.

The framework comprises a workflow with 4 stages. The initial stage is data gathering and processing. It covers the activities of identifying, selecting and preprocessing incident related data. It is a feature selection and construction process. Next, a set of machine learning algorithms were selected and applied over the set of selected features, therefore descriptive models were built. We follow the definition proposed by Meneses (Meneses and Grinstein 1998) where description concerns in finding patterns that describe the model represented by the data or even the process generating the data. Following, an evaluation process was performed, where measures of performance accuracy were computed to identify the models with better predictive performance. The last stage of the workflow relates to the usability of the descriptive models, where a visual analytic approach was chosen. Different visualizations techniques have been investigated; two of them were initially chosen and implemented as part of this computational framework. Figure 1 illustrates this workflow.

Figure 1. Analytic framework’s workflow

Data Gathering and Processing

Besides the challenge of identifying a computing solution for incident data analysis and prediction, there is the challenge of identifying pieces of data that are relevant for such analysis. For example, Evans (Evans 2011) identified the following as broad cause of railway incidents: signal passed at danger, overspeeding, signaling or dispatching error, other operational error, rolling stock failure, problems in the infrastructure track or point failure and external events to railway. Lira (Lira et al. 2014), besides operational data, also took into account socio economic data.

The project described in this document uses operational data provided by Vale’s incident reporting system. We have accessed operational data from a period of 9 years, with fifty-nine attributes (features), and 27282 instances. Therefore, the initial work was to get an understanding about the dataset, and the meaning of its features. Vale’s safety engineers help us to understand the meaning of dataset’s features, as well as the complexities of the domain (railways incident analysis). Therefore, features selection was initially based on experts (safety engineers) information. Based on that understanding, irrelevant features were identified and removed. For example, many features had an administrative purpose such as to store
the time and date a report was issued (when incident's data was entered in the system), others feature related to financial costs and operational loss incurred as a consequence of an incident. Although those can be considered important information, they have no value for predictive purposes, and were removed. Others feature were removed based on analysis about their frequency distributions; highly sparse features were removed.

To illustrate, some of the features used in our investigation were: the time and date when an incident occurred, the municipality, the position along the railway (given in kilometers) where an incident occurred and the existence of level crossings. We have also classified the convoys according to the number of locomotives (a convoy can have up to 4 locomotives), according to the number of loaded wagons and empty wagons, and according to the length of a convoy (a convoy can be up to 3 kilometers long).

Among the initial 59 features, one has been especially important, “Natureza”, with 76 distinct possible values. Natureza identifies a particular type of incident, for example, DESCAR for derailment, COLISA for collision, DEFEE for failure in an electric equipment, and so on.

Among the 76 types of incidents, our initial effort focused on two types of incidents and later was extended to other types, which were indicated to us by Vale's railways safety engineers, as the most critical for analysis and prevention. For the reason of data confidentiality, we have labeled these incidents as INCA, INCB and INCC respectively. Therefore, it becomes a classification problem, with 3 classes, where class INCA was the most important (critical to prevent), followed by INCB.

There were continuous attributes that had to be discretized for the purpose of data mining, such as the total length of a convoy (which is given in meters), the total weight in tons, and the kilometer along the railway where an incident occurred, among others. Some attributes have been created, as the visibility index over the railway. These are briefly discussed to exemplify the work that has been done, but we cannot possibly include in this document all the data processing activity that has been developed.

The railway in our study has nearly one thousand kilometers of extension, and the data associated with the location of an incident over the railway is given in kilometers plus meters, such as 370.182 or 441.58. Table 1 illustrates some cases of incidents and respective position over the railway. For example, the first line in Table 1 indicates that an incident type INCC occurred at the kilometer 398, plus 148 meters.

<table>
<thead>
<tr>
<th>Kilometers</th>
<th>Incident</th>
</tr>
</thead>
<tbody>
<tr>
<td>398.148</td>
<td>INCC</td>
</tr>
<tr>
<td>260.814</td>
<td>INCA</td>
</tr>
<tr>
<td>341.282</td>
<td>INCA</td>
</tr>
</tbody>
</table>

Table 1. Incident positions along the railway

An equal-frequency discretization technique based on the INCA type of incident (class) has been applied to discretize the kilometer into ranges of kilometers. Equal-frequency discretization is an unsupervised binning method, where data is divided into groups (bins). Each group contains approximately same number of values. As a result of this discretization procedure, a categorical attribute with eight (8) intervals along the railway has been computed. The aim is to capture parts of the railway where incidents have occurred in a level of granularity with enough descriptive significance. There isn’t a rule that gives an optimal number of intervals (bins), therefore, after an investigation, we chose to use eight intervals, as this allows us to better represent the railway in ranges, according to the occurrence of the incidents under study.

The visibility index is a measure of the size, or length, of the viewed area taken from the train driver position in a certain point over the railway. This is an important aspect to be taken into account when studying incidents in railways since the terrain topology influences the driver’s viewed area. The visibility index is a way to account for this. It is calculated through the elevation value of a digital elevation model (DEM) (Lee and Stucky 1998) of the railway. Figure 2 illustrates the concept of a visibility index. A human figure is placed in a position over a pathway. P1, P2 and P3 are locations over this pathway, and the lines indicate the visibility. In Figure 2, P3 and P1 are fully visible from the position of the human figure, P2 is not visible as there is an elevation preventing its visibility. The dashed line indicates the limit positions of visibility.
In our study we computed the train driver visibility in meters, taken at a 1 kilometer backwards from a point where an incident occurred. The purpose is to investigate whether the driver's visibility has an impact on incidents, and therefore could be used as predictor. In a similar fashion as the kilometer attribute, the visibility index was calculated in meters, and transformed into a categorical attribute.

Attributes selection was primarily done based on domain knowledge. After irrelevant attributes have been identified and removed, a feature selection algorithm was executed to evaluate the discriminative power of each of the remaining attributes. The Gini Index algorithm implemented in the Weka system, version 3.8 (Hall et al. 2009), was the algorithm applied for this evaluation. The final dataset for analysis has 17 attributes.

**Descriptive Models**

In this stage, a set of descriptive models has been built. As already mentioned, we follow the definition proposed by Meneses (Meneses and Grinstein 1998) where description is concerned with the activity of finding patterns that describe the model represented by the data, therefore, we will be looking for patterns that describe (railway) incidents.

The term **analytical dataset** has been used to distinguish a dataset which has an analytic purpose, as oppose to an operational purpose. The basic difference is on how the data is defined and structured. For example, an operational dataset typically would have a timestamp data, which identifies a time when an event happened, an incident for instance. For analysis purposes, a timestamp is of little value; instead, an analytical dataset would record a particular day of the week, a month, or a period along the year (summer, spring, fall or winter for instance). This kind of data allows for finding patterns within the dataset.

The analytical dataset in our study contains 17 attributes, and 27282 records, distributed across all incident types. Figure 3 illustrates the initial class distribution within the population.

It can be observed in Figure 3 that classes in our dataset are clearly unbalanced. The vast majority of the population, 70.59%, belongs to class INCC, and the classes of interest, INCA and INCB, represent 20.59% and 8.82% respectively. This unbalanced class distribution imposes a problem in building classifiers (Viademonte and Burstein 2006; Aggarwal 2015).

![Figure 2. Visibility index](image)

A random sampling approach was applied to achieve a more homogeneous class distribution; the dataset with class INCC was randomly sampled on 30% of its initial population, to reduce its size (under sampling). Analytic datasets are used as input parameters for machine learning algorithms. As we are concerned in identifying classes (types) of incidents, the results of this computation are termed "classifiers".

This is an exploratory analysis, the choice of which machine learning algorithm to apply was part of the investigation. Therefore a set of algorithms that are often applied for classifications purposes were chosen: a decision tree learner algorithm, logistic regression, support vector machine (SVN), and Naive
Bayes. The implementation of those algorithms in WEKA 3.8 (Hall et al. 2009) environment and Orange 2.7 (Demsar et al. 2013) were used. Next section discusses the evaluation of those experiments.

**Evaluation**

There are different approaches to evaluate a classifier on a given dataset, such as the holdout method, cross validation and bootstrap (Aggarwal 2015; Hastie et al. 2009). Typically, a dataset is divided in two sets, one with labeled data for learning (supervised learning), and another with unlabeled data for testing purposes. Learning sets are selected on higher proportions than test sets, usually between 70% and 80% of the whole dataset. The remaining data is selected for test purposes. This study applied a 10 fold cross validation procedure. In cross validation, the data set is divided into $k$ disjoint subsets of equal size, in this case $k = 10$. One of the k subsets is used for testing, and the others subsets for learning. This approach is repeated by selecting all of the $k$ subsets in dataset (Aggarwal 2015).

The measures of *classification accuracy*, *precision* and the *ROC curve (AUC)* were used to estimate the classification performance of the classifiers. *Classification accuracy* is the fraction of test instances correctly classified, for example, an observed instance of an incident INCA correctly classified as INCA. *Precision* is a measure that is suitable for cases with unbalanced class distributions, and it is the ratio between true positives and the sum of false positives with true positives. The *AUC* gives a performance measure of a classifier by plotting true positive rates against false positive rates. Detailed discussions about measures of accuracy for classifiers, or predictive models in general, can be found in (Aggarwal 2015; Hastie et al. 2009; Vapnik 2000; Wagstaff 2012). Figure 4 illustrates the performance measures of the classifiers in our research.

![Figure 4](image)

**Figure 4. Performance measures for the classification of incidents**

The algorithm with best classification performance was logistic regression, with a classification accuracy of 0.9375, followed by Naive Bayes with 0.9276. However, those experiments are not conclusive, as new instances are added in the data set, and some of the decision we took related to data modeling may change according to feedback gathered from our users, safety engineers. For example, the ranges used to discretize some continuous attribute, such as kilometers, may change. In general, data sets and their features are prone to change over time, and as a consequence, the models (classifier).

**Visual Analytic Approach for Knowledge Exploration**

One of the challenges in applying predictive models to decision support is on how to deploy such models to users and operations. To apply a normative approach, i.e., to present to users the results of a computation, such as a probability, without an explanation on how this computation has been performed, or which pieces of information it has applied, is likely to compromise the acceptance of such (predictive) models. There are many studies and initiatives in the subject of providing explanations for the computation of machine learning algorithms, such as the LIME approach (Ribeiro et al. 2016).

Users of our application told us they would like to be able to visualize the information associated to the computation of the risk index, i.e., they would like to know which conditions contributed to the computation of such index. They expect to obtain new insights about safety conditions that impact operations. This brings additional challenges, as the computation of machine learning algorithms are, usually, not transparent (Martens and Provost 2014; Ribeiro et al. 2016). Therefore, we decided to create a visual analytic capability that would expose the knowledge computed by a classifier: a knowledge visualization capability.

We have decided to use a decision tree algorithm for this purpose. The reasons decision trees were chosen are because it has an explicit and transparent way of representing data, in a hierarchical fashion. Also, decision trees showed an acceptable level of performance in the evaluation phase.
Several visualization techniques have been analyzed, through the D3 gallery (D3 Gallery 2017). The first visualization technique that was chosen was the Sunburst. The Sunburst technique was initially chosen as it nicely represents data hierarchies and relations. Sunburst allows users to navigate through correlated sets of features in an intuitive and exploratory fashion, therefore suitable for our purposes.

Sunburst represents hierarchies through a series of rings. Each ring corresponds to a level in the hierarchy, and the central circle in the Sunburst represents a root (top) node in a tree structure. External rings in the Sunburst correspond to leaf nodes in a tree structure.

Figure 5 illustrates how the Sunburst has been applied to implement knowledge visualization in the project described in this document. In Figure 5, railway incidents (classes) are represented by colors, red color would represent incident INCA, yellow incident INCB and green INCC. The color’s intensity of a particular path in the Sunburst represents the strength of the relationship represented by that path. The circle in the middle in the Sunburst represents the root node of a decision tree, and the nodes at the edge of the Sunburst represent leaf nodes.

![Figure 5](image.png)

**Figure 5. Sunburst technique to visualize and browse a decision tree model**

Therefore, navigating from the central node towards an edge node in the Sunburst is equivalent of moving downwards in a decision tree, from the top node towards a leaf node. In the computational framework implemented as part of this research project, nodes in the Sunburst are all clickable. Users can drill-down the Sunburst according to the level of information they are interested, by clicking on a particular node.

Figure 5 illustrates this feature. The surrounded area in the Sunburst represents the following association: type = Circulation, Kilometers = 600-740, traction = conventional, month = January, and period of the day = night (readers should note this is the English translation of the respective elements which are originally in Portuguese). This association is related to an incident INCA (red color at the leaf node) with confidence level of 0.619. There is a left panel in Figure 5, containing the decision tree elements.

At the top of the left panel, there is a slider bar. This bar allows user to filter relations according to a threshold, e.g. a confidence level. Below the slider bar, incidents and respective colors are displayed. At the bottom of the panel, associations that have been selected in the Sunburst are displayed. The visualization capability also allows users to select incidents they are more interested in.

Besides the Sunburst, we were also interested in providing a visual capability that allows a top-down representation of the computed knowledge. For this, the Circle Packing technique in D3 gallery (D3 Gallery 2017) was selected. Circle Packing represents nodes of a tree structure as circles, and nodes belonging to subtrees, are nested inside the circles. Figure 6 illustrates the Circle Packing representation of a decision tree, in our study.

Similar to the Sunburst, incidents are also represented by colors in the Circle Packing. The intensity of colors indicates the strength of an association. The right side panel in figure 6 enumerates the incidents in our study (INCA, INCB and INCC). The big light blue background represents the top node in the decision tree, and the two large inner circles, red and green, represent the nodes in the level below. The incidents
to which these nodes are associated with, are given by their respective colors. Clicking in a circle, will take users to the next level down in the tree structure, until the circles that correspond to leaf nodes are reached.

The Circle Packing, as the Sunburst, is a fully clickable feature. Clicking in a particular node will highlight the subtrees structure under that node, and clicking outside a node, will move up in the tree structure (backwards). The navigation view, incidents and their respective colors, and associations corresponding to the selected decision tree node are all highlighted at the right side panel.

Figure 6 illustrates the implementation of the Circle Packing technique in our research.

![Figure 6. Circle Packing technique to visualize and browse a decision tree model](image)

This visual analytic feature has been presented and discussed with users, Vale’s employees. The feedback obtained from the users has been positive and valuable, not only to assess the usefulness of the application as whole, but also the suitability of this visualization approach. The capability of selecting a particular incident and explore the conditions that lead to its occurrence, showed to be very helpful in assessing the conditions that contribute to the occurrence of such incident.

**The Consulting Capability**

The project described in this document aims to deploy predictive models as an advisory computer system, to support the tasks of railways incident management and risk assessment. Therefore, the computational framework has the capability of querying a (predictive) model. Presented with a set of input data, it computes an estimative about the occurrence of incidents. At the time this document has been written, a decision tree algorithm within the WEKA API (Hall et al. 2009) has been used to implement the predictive model component.

The consulting capability presents an interface with a form for data input and a topographic map of the railway under study. The form is where users select the characteristics of a particular convoy. Figure 7 illustrates this feature. The left side panel in figure 7 displays the characteristics of a particular convoy, for example: the type of a convoy (this is related to the kind of load a convoy is transporting, such as minerals, grains, passengers, etc.), the length of a convoy (convoys can be as long as 3 kilometers), the number of fully loaded and empty wagons and the number of locomotives. It also presents time and location related information, such as month and day of the week, if a convoy is circulating at night, morning or afternoon, if there is a level crossing in a particular location, among other characteristics.

The input data is presented to the predictive model component for assessment, i.e. is passed as input arguments. The computation occurs as users move the mouse along the railway track. The computed value, risk index, is highlighted for that particular area over the railway. For example, figure 7 illustrates the computation of risk index in the range between kilometers 310 and 460. In this example, given the characteristics at the right side panel, on that position over the railway, the highest risk index is 0.44 for incident INCA, followed by 0.30 and 0.26 for incidents INCC and INCB, respectively.
Conclusion

This document presents a research project related to the development of a computational framework to support the tasks of railways incident analysis and risk assessment. The computational framework applies data mining, machine learning and data visualization technologies, to generate knowledge from incident data and allows users to visually explore such knowledge, to obtain insights and awareness about conditions that impact railways safety.

The contributions of the project discussed in this document are the development of a computer application, using data mining, machine learning and data visualization technologies. The data visualization component exposes the domain knowledge computed by a machine learning algorithm, allowing users to visualize and explore it. The experience of the development of this computational model for railways incident analysis, the study of feasibility of such approach for railways operations, and the identification of relevant information for this specific type of analysis.

The stages of data processing show us the need to come up with a more dynamic approach for data gathering and processing. Railways data are surprisingly dynamic. Railways level crossings are built, new tracks can be deployed along railways, communities settle and evolve around railways, and those changes, often, need to be included in the data model for data mining and predictive purposes.

The majority of the data used in this project relates to operational data, as this is the kind of data that is more readily available and organized, but other sources of data can be also relevant. Sensor measurements taken from different devices placed along the railway should be considered, as well as locomotives and wagons maintenance data, to enhance the analytic models. Those different sources of data will be considered in future developments of this research.

In this research we clearly distinguish data from knowledge. Where knowledge is the output of a particular machine learning algorithm, and data is the input of such algorithm. Sunburst and Circle Packing techniques have been implemented to allow users interactively visualize knowledge (railways incident). A challenging research topic is on how to turn explicit the knowledge produced by machine learning algorithms. This is particularly challenging for algorithms such as support vector machines, where domain knowledge is implicitly encoded within its computation. Knowledge visualization is a research topic that will be further investigated as part of the research described in this document.

Acknowledgements

The authors would like to acknowledge the contribution and support from the staff of Vale's Carajás railways. This research has been partially funded in cooperation between the Instituto Tecnológico Vale and the Brazilian National Council for Research and Development (CNPq), under the research grants 440880/2013-0 and 310468/2014-0.
REFERENCES


