Proceedings of the 2018 Joint Rail Conference
JRC2018
April 18-20, 2018, Pittsburgh, PA, USA

JRC2018-6242

NOVEL INSIGHTS FOR RAILROAD MAINTENANCE USING BIG DATA ANALYTICS

Naji Albakay
University of Nebraska-Lincoln
Omaha, NE, USA
nalbakay2@unl.edu

Michael Hempel
University of Nebraska-Lincoln
Omaha, NE, USA

Hamid Sharif
University of Nebraska-Lincoln
Omaha, NE, USA

ABSTRACT

Train accidents can be attributed to human factors, equipment factors, track factors, signaling factors, and Miscellaneous factors. Not only have these accidents caused damages to railroad infrastructure and train equipment leading to excessive maintenance and repair costs, but some of these have also resulted in injuries and loss of lives. Big Data Analytics techniques can be utilized to provide insights into possible accident causes, thus resulting in improving railroad safety and reducing overall maintenance expenses as well as spotting trends and areas of operational improvements. We propose a comprehensive Big Data approach that provides novel insights into the causes of train accidents and find patterns that led to their occurrence. The approach utilizes a combination of Big Data algorithms to analyze a wide variety of data sources available to the railroads, and is being demonstrated using the FRA train accidents/incidents database to identify factors that highly contribute to accidents occurring over the past years. The most important contributing factors are then analyzed by means of association mining analysis to find relationships between the cause of accidents and other input variables. Applying our analysis approach to FRA accident report datasets we found that railroad accidents are correlating strongly with the track type, train type, and train area of operation. We utilize the proposed approach to identify patterns that would lead to occurrence of train accidents. The results obtained using the proposed algorithm are compatible with the ones obtained from manual descriptive analysis techniques.

INTRODUCTION

The North American railroad industry generates significant amounts of information related to its operational efficiency [1-4]. This information includes operational data, accidents/incidents data, track maintenance data, safety data, inventory and highway-rail crossing data, and inspection and maintenance data [5]. Traditionally, these data sets have been stored in multiple databases and analyzed independently using traditional descriptive analysis techniques. However, these databases can be brought together and analyzed using Big Data Analytics techniques in order to uncover hidden patterns and find correlations that might not be easily discovered from analyzing data separately. In addition, Big Data analysis would allow the usage of predictive and perspective analysis techniques to forecast future safety measures and provide insights into possible accident causes, manufacturer issues, and more. For instance, Big Data Analytics tools can combine railroad accidents/incidents database with operational and maintenance databases and allow for prediction of train failures before they occur. It could also allow for efficient scheduling of train and track maintenance thus enhance rail safety and reduce the costs caused by unnecessary maintenance. There are predictive Big Data algorithms that are well known for their accuracy including Random Forest (RF) and association mining algorithm. RF is the most popular algorithm in conducting in-depth study of Big Data [6]. It has classification and regression capabilities and high-performance efficiency. RF also gives estimates of what variables in the input data are more important in achieving certain responses [7]. This latter property is very significant as it enables selecting the important features and build a simple model based on these features, thereby reducing the computational cost. Association mining algorithms, on the other hand, analyze the input data set for frequent patterns [8]. They automatically find the patterns that would take a long time to find manually using descriptive analysis techniques. The advantage of association algorithms over RF algorithms is that associations can exist between any of the input variables. While the RF algorithm

builds rules with only a single conclusion, the association algorithms attempt to find many rules, each of which may have a different conclusion. Association algorithms use the support and confidence criteria to identify the most important relationships. Support is an expression of how frequently the variables appear in the input data, whereas confidence expresses how often that relationship has been found to be true within the data set. The main drawback in association algorithms is the computational efficiency as they require extensive processing time to find patterns within a potentially large search space [9-11].

In this work we develop a comprehensive Big Data algorithm that utilizes the importance measurement feature from RF algorithm and the pattern detection capability of association mining algorithms. The importance measure helps in choosing the most important variables in the input data and thus increase the computation speed of the association mining algorithms. The remainder of this paper describes the algorithm structure and the results.

METHODOLOGY

The proposed algorithm utilizes both RF and association mining algorithms. RF allows selection of the most important variables in the input data subject to a specific response and feeds them to the association algorithm that discovers the connection between the variables. Here is the algorithm pseudocode:

1. Let N be the number of rows in the input data, M be the number of columns and K is a subset of the possible categories
2. Determine $m \subseteq M$ such that m has high impact on deciding K, using the importance feature from RF algorithm
3. Find $X \rightarrow Y$ where $X \subseteq m$, $Y \in K$ and $X \cap Y = \emptyset$
4. Find support $\sigma(X \rightarrow Y)$ and the confidence $C(X \rightarrow Y)$[11]
5. Choose $X \rightarrow Y \ni \sigma(X \rightarrow Y) > 0.1$ and $C(X \rightarrow Y) \geq 0.8$

RF used in step 2 is an aggregation of decision trees where every node in the tree is used as a binary condition on a single variable of the input data set. The condition at each node splits the variables into two groups, such that each group contains data that provides a similar response. The measure of the optimal splitting condition is based on Gini impurity. When training RF with the input data set, the decrease in the weighted impurity caused by each variable of the input data set is computed. The impurity reduction caused by each variable is averaged and the variables are ranked according to this measure. Variables that can remove more impurity are ranked as more important than the ones that remove less impurity. We can think of the important variables (m) as the ones who contributed the most to the rules formed by RF algorithm and thus a change in their value would degrade RF prediction ability as measured by out-of-bag (OOB) techniques [11].

The implication relationship in step 3 is the association mining rule where X and Y are called antecedent and consequent, respectively. In step 4 we select the rules from the set of all possible rules found by the association mining algorithm constraints to the thresholds on support and confidence measures. A rule is identified as important if the confidence and the support are within 0.8 and 0.1, respectively.

IMPLEMENTATION AND RESULTS

A. INPUT DATA

The proposed algorithm was implemented in RStudio [12] by leveraging both RF and "arules" packages. In order to assess the algorithm efficiency, we tested the algorithm on the Federal Railroad Administration (FRA) accident/incidents database and compared the obtained results with the ones from manual descriptive analysis.

The input data set used is from the Federal Railroad Administration accident data sets [13] obtained for the period from January 2013 to December 2016. It contains information regarding a variety of conditions or circumstances that may have contributed to the occurrence of the reported accidents. The data accounts for damages to on-track equipment, signals, track, track structures, and roadbed. It comprises 50 columns (M), which are the fields from the "F.6180.54" form, and 9864 rows (N) that represent the number of accident/incident reports filed over the mentioned time period. According to the data base, there are five major classes (K) of train accidents, namely: human factors (H), equipment factors (E), track factors (T), signaling factors (S), and miscellaneous factors (M). The number of accidents in each accident cause category is shown in Fig. 1.

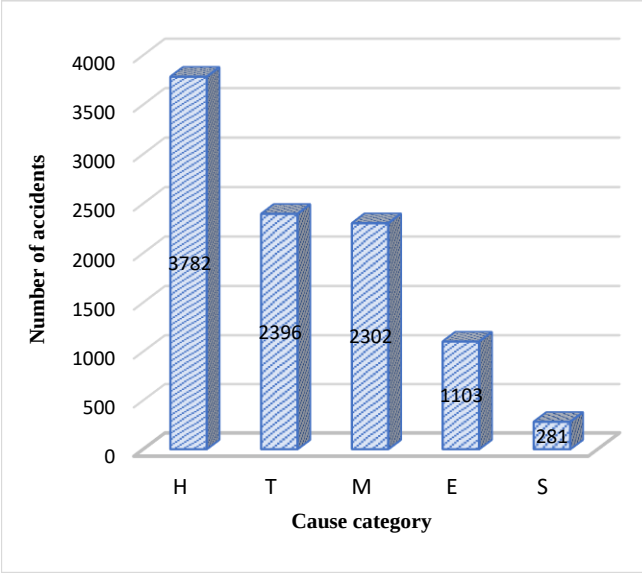

Fig. 1: Accident causes category versus number of accidents

B. IMPORTANT FEATURES SELECTION

The input data is applied to the RF algorithm in order to find the variables that contributed the most to the cause of these accidents, based on the mean decrease in Gini impurity.

Fig. 2 displays the 30 most important variables in the input data on the y-axis and the mean decrease in Gini score on the x-axis. A higher value of mean decrease in Gini score implies a higher importance of the associated variable. For example, the grade crossing ID number (GXID) and the DRUG in Fig. 2 are the most important variables in predicting the cause of accident. Table 1 lists the most important variable and their description.

Table. 1: Most Important Variables

Feature acronym	Description
GXID	Grade crossing ID number: 0= No grade crossing, 1= Grade crossing
DRUG	Number of positive drug tests: 0=No positive drug test reported, 1=positive drug test reported
Longitude	Longitude in decimal degrees
TRKNAME	Track name
Latitude	Latitude in decimal degrees
ALCOHOL	Number of positive alcohol tests 0=No positive alcohol test reported, 1=positive alcohol test reported
HIGHSPD	Maximum speed reported for equipment involved
STATION	Nearest city and town
RRCAR1	Car initials (first involved)
STCNTY	FIPS State & County code
TEMP	Temperature in degrees Fahrenheit
TRNNBR	Train ID number
LOADF1	Number of loaded freight cars
TRNSPD	Speed of train in miles per hour
Column1	Gross tonnage, excluding power units
TIMEHR	Hour of incident
IMO	Month of incident
STATE	FIPS State code
EMPTYF1	Number of empty freight cars
RAILROAD	Reporting railroad
TRKDNSTY	Annual track density - gross tonnage in millions
LOADED1	car loaded or not (first involved): Y=yes N=no blank=not applicable
TYPEQ	Type of train: 1=freight train, 2=passenger train, 3=commuter train, 4=work train, 5=single car, 6= cut of cars, 7= yard / switching, 8= light loco(s), 9= maintenance / inspection car
ENGHR	Number of hours engineers on duty: blank=not applicable
TYPTRK	Type of track: 1=main, 2=yard, 3=siding, 4=industry
CDTRHR	Number of hours conductors on duty: blank=not applicable
HEADEND1	Number of head end locomotives

The most important variables are applied to the association algorithm, which resulted in 58987 patterns. However, it is clear that going through all these patterns manually is not a viable option. Therefore, we used the scatter plot to visually see the rules and interactively choose the most significant ones based on their confidence value. The scatter plot of the confidence and the

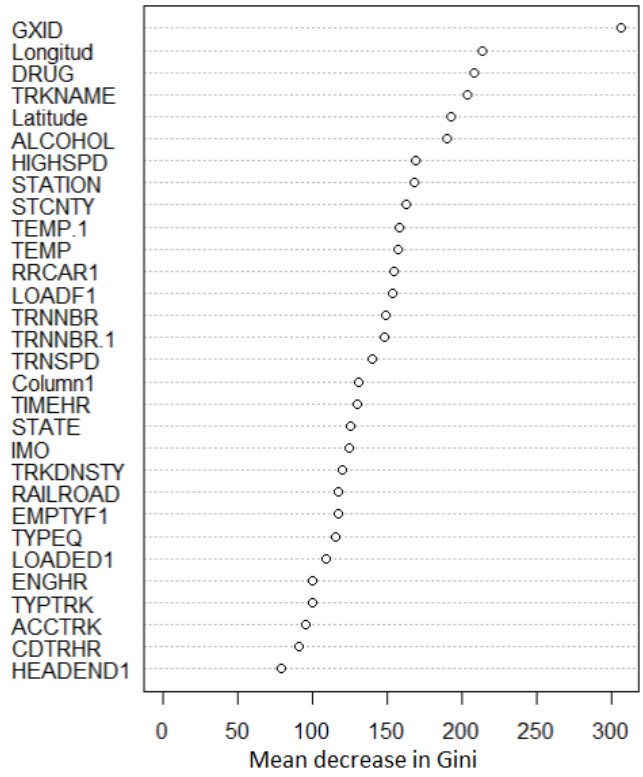

Fig. 2. Importance plot

support for all rules is shown in Fig. 3. The plot consists of the support as x-axis and confidence as y-axis and each dot on the

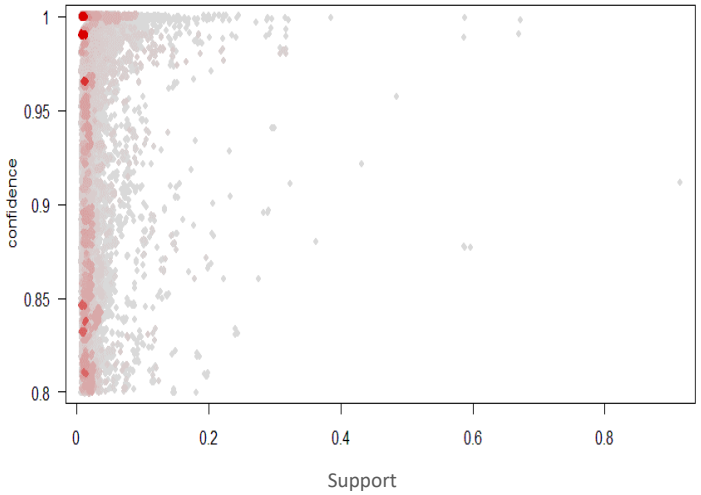

Fig. 3: Scatter plot

plot represents one of the obtained rules. We adjust the logarithm so that we can see only the patterns with confidence higher than 80%. Also, the dots are color coded so that the red dots indicate that the rule has high confidence value and needs to be further explored.

Table 2 displays an example of the four most important patterns and their support and confidence. These patterns are regarded as important because the confidence is above 80%. The first rule states that accidents due to human factors (H) often occur at rail yards (TYPEQ= yard / switching) when no grade crossing is involved (GXID=No) and the train engineers are not under the influence of drugs (DRUG=No). The algorithm also states that this pattern is 98.3% reliable and applies to 12.4% of the input data. By analyzing the data manually, we found that among the 3782 accidents that are caused by human errors, 1397 accident occurred at the rail yards when train engineers tested negative on drugs and no *GXID* report was found. Therefore, the manual results are compatible with the automatic results obtained using the proposed algorithm. The second pattern states that the accidents caused by Miscellaneous factors (M) often occurs to passenger trains (TYPEQ=Passenger train) on a single main track (TRKNAME=Single main track) when train engineers are not on drugs (*DRUG=No)*. It also states that this pattern applies to 10.5% of the input data and has 97.8% reliability. Manual analysis confirms that the highest number of accidents (34/35) due to Miscellaneous (M) factors occurred to passenger train on a single main track when train engineers tested negative for drugs. The third significant pattern in Table 2 implies that accidents caused by track factors (T) often occurs to freight trains (TYPEQ= Freight train) in state 48 (Texas) given no alcohol (ALCOHOL=No) or drugs (DRUG=No) are involved and no GXID (GXID=No) is involved. It also states that this pattern applies to 19.5% of the input data and has 94.4% reliability. This also agrees with the manual analysis which show that among the 639 accidents that happened to freight train in Texas, 200 accidents occurred due to track factors as illustrated in Fig. 4.

Table. 2: Sample of the most important patterns

Rule	Support	Confidence
{GXID=No, DRUG=No, TYPEQ= yard / switching} => {CAUSE=H}	0.124	0.983
{DRUG=No, TRKNAME=Single main track, TYPEQ=Passenger train} => {CAUSE=M}	0.105	0.978
{GXID=No, DRUG=No, ALCOHOL=No, TYPEQ= Freight train, State=48} => {CAUSE=T}	0.195	0.944
{DRUG=No, TYPEQ= Freight train, State=48 } => {CAUSE=E}	0.165	0.912

The last rule implies that most accidents caused by equipment factors (E) are occurring for freight trains (TYPEQ= Freight train) in state 48 (Texas) when the engineers are tested negative for drugs (DRUG=No). It also states that this pattern applies to 16.5% of the input data and has 91.2% reliability. This also agrees with the manual analysis, which shows that among the 639 accidents of freight train in Texas, 58 accidents occurred due to equipment factors as illustrated in Fig. 4.

Notice that the algorithm can be used to predict what cause category the accident should fall into. For instance, given that an accident occurred at rail yards (TYPEQ= yard / switching) when no grade crossing is involved (GXID=No) and the train engineers are not under the influence of drugs (DRUG=No), according to the first rule in Table 2, we can predict with 98.3% accuracy that the accident is caused by human factors.

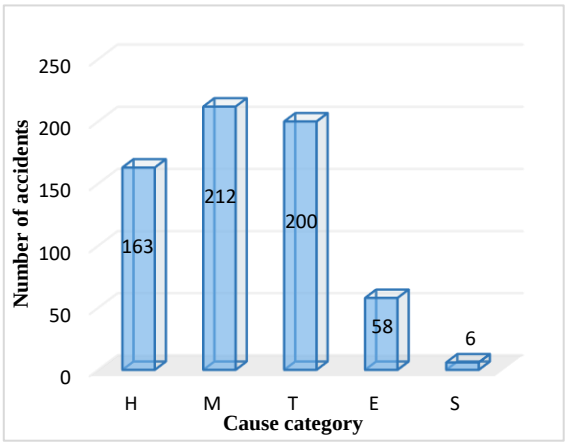

Fig. 4: Freight train accidents in Texas between 2013 and 2016.

DISCUSSION

Studying more closely the rules in Table 2 we can observe that the variables that appear in the rules are the ones that have the largest mean Gini score shown in Fig. 2. The variable DRUG, for instance, is associated with every accident because the actual number of accidents involving DRUG is zero. GXID, on the other hand, has such a large mean Gini score since grade crossing accidents are quite common and the presence of a grade crossing is always a factor for a crossing related accident. Notice that all the variable that are regarded as important according to Table 1 have appeared in some rules generated from the association mining. However, some of those rules might have low confidence.

One of the main advantages of the proposed algorithm as compared to the manual analysis is the ability to detect and extract useful information from large-scale data with high computational speed and is scalable to very large datasets not feasible for manual analysis. Another key differentiator is that with the proposed approach it is possible to detect the impact from weaker correlations among different parameters that may not be apparent using manual analysis.

The proposed algorithm has predictive capabilities since it utilizes Association mining and RF algorithms, both of which are predictive algorithms. However, maintenance and operational data sets are needed for these predictive capabilities. The ultimate goal of our work is to apply the algorithm for predictive analysis, so we can reliably predict what cause category the accident should fall into and therefore help the railroads and government agencies improve the productivity, reliability, and safety of their operations. We are also working on applying the proposed algorithm to an integrated database that comprises operational data, accidents/incidents data, track maintenance data, safety data, inventory data, highway-rail crossing data, and inspection and maintenance data. The goal is to process the data for new insights into component failure prediction, maintenance schedule optimization, replacement component selection, and failure cause analysis.

CONCLUSION

We proposed a comprehensive Big Data algorithm that utilizes the importance feature from RF algorithms and the pattern detection ability from the association mining algorithms in combination to reduce computational complexity while retaining all the insights available from the collective data set. The developed algorithm was applied to the FRA accidents/incidents data as an evaluation tool for its efficacy and has shown results similar to the results that were obtained using manual analysis, thus validating its accuracy. Our work shows that Big Data analytics applies to maintenance and operational data can reliably identify accident categories and cause factors, and thus assist with improving the productivity, reliability, and safety of the rail operations.

ACKNOWLEDGMENT

This study was conducted at the University of Nebraska-Lincoln by the research faculty and students at the Advanced Telecommunications Engineering Laboratory (www.TEL.unl.edu). This project is supported by the University Transportation Center for Railway Safety (UTCRS).

REFERENCES

[1]. M. Zarembski, "Some examples of big data in railroad engineering", 2014 IEEE International Conference on Big Data (Big Data), 2014.

[2]. A. Zarembski, "Integration of multiple inspection system data to identify potentially unsafe track rail conditions: data collection, consolidation and preparation", Report Prepared for US Federal Railroad Administration, 2014.

[3]. N. Attoh-Okine, "Big data challenges in railway engineering," IEEE International Conference on Big Data (Big Data), 2014.

[4]. D. Hunt, J. Kuehn, and O. Wyman, "Big data and rail- road analytics". Newsletter of the Railway Applications Section, 2013.

[5]. Federal Railroad Administration(FRA): https://www.fra.dot.gov/Page/P0001

[6]. Y. Xu, "Research and implementation of improved random forest algorithm based on Spark", IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017.

[7]. W. Lin, Z. Wu, L. Lin, A. Wen and J. Li, "An ensemble Random Forest Algorithm for insurance big data analysis", IEEE Access, 2017.

[8]. A. Behnamian, K. Millard, S. N. Banks, L. White, M. Richardson and J. Pasher, "A systematic approach for variable selection with Random Forests: achieving stable variable importance values", IEEE Geoscience and Remote Sensing Letters, 2017.

[9]. Y. Motai, "Kernel association for classification and prediction: a survey", IEEE Transactions on Neural Networks and Learning Systems, 2015.

[10]. H. Yan and H. Hu, "A Study on association algorithm of smart campus mining platform based on big data", International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2017.

[11]. K. Max, and K. Johnson, "Applied predictive modeling", Vol. 26, Springer, 2013.

[12]. RStudio homepage, [Online] https://www.rstudio.com/

[13]. Federal Railroad Administration Office of Safety Analysis, [Online]http://safetydata.fra.dot.gov/OfficeofSafety/Default.aspx