Leveraging Spark and Docker for Scalable, Reproducible Analysis of Railroad Defects

Dylan Chapp
University of Delaware
dchapp@udel.edu

Surya Kasturi
University of Delaware
suryak@udel.edu

ABSTRACT

1. INTRODUCTION

According to the United States Federal Railroad Administration Office of Safety Analysis, track defects are the second leading cause of accidents on railways in the United States. In light of the economic significance of railway accidents [1], there is a pressing need in the railroad engineering community to adopt data-driven scalable data analysis tools from the greater "Big Data" ecosystem. [3] Track maintenance–i.e., identifying and repairing defects–is one of the primary factors that affect the service life of a rail track, but due to the severe safety implications of undeftected or unrepaired defects, the ability to predict common defects is highly desirable.

In this work, we present a case study centered on the analysis of two railroad defect data sets obtained from railroad engineering researchers in the University of Delaware Department of Civil Engineering. Hereafter we will refer to these datasets as the `rail_defects` data set and the `track_geometry_defects` data set. Respectively, these data sets describe defects in the rails themselves, such as voids or internal changes in crystalline structure, and misalignment of track components, such as one rail tilting away from the other. We investigate the feasibility of predicting the type of a defect based on associated data such as geographic region, mean gross tonnage (MGT) the track is subject to, and rail type. In the rest of this paper, we outline the construction of our analysis platform, present some initial results on classification accuracy, and propose extensions to our work.

2. METHODOLOGY

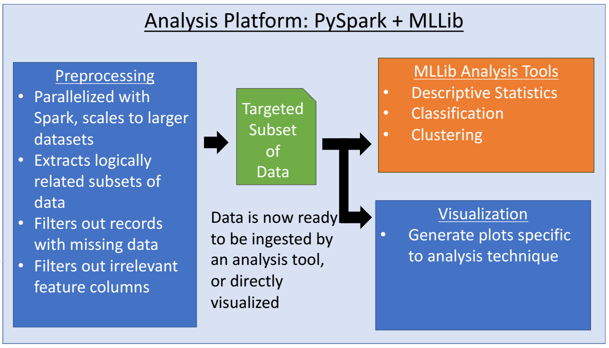
The data sets under consideration present familiar, but nevertheless nontrivial challenges. The `rail_defects` data set consists of 26,432 20-dimensional points, and the `track_geomet` data set consists of 25,421 41-dimensional points. Both data sets contain mixed numerical and categorical data, and each data set's points are each labeled with a class label indicating the kind of defect the point refers to.

In light of the properties of the `rail_defects` and `track_geometry_defe` data sets, we initially focus on obtaining accurate predictions from multilabel classification models. We argue that studying the accuracy of standard multilabel classifiers against our static data sets is a crucial first step towards models that can predict likely defects for a particular section of railroad.

With a view towards coping with larger data sets than those currently under consideration, we implement our analysis plaform using the MapReduce framework Apache Spark. Spark provides improved performance for the data transformation tasks we perform during preprocessing relative to other MapReduce frameworks [2], and offers convenient Python bindings. Moreover, Spark comes packaged with a full-fledged paralllel machine learning library–MLLib–whose classifier implementations we use.

In order to guarantee portability and reproducibiliy of our analyses, we package our code, data, and environment as a Docker container. In addition to aiding our own development effort by enabling rapid, safe iteration on our design, containerization allows collaborators who are unfamiliar with Spark or MLLib to run our code without doing any configuration of their environment beyond setting up Docker.

We present a block diagram of our platform's high-level design in Figure 2


Analysis Platform: PySpark + MLLib
Preprocessing
Parallelized with Spark, scales to larger datasets
Extracts logically related subsets of data
Filters out records with missing data
Filters out irrelevant feature columns
Targeted Subset of Data
Data is now ready to be ingested by an analysis tool, or directly visualized
MLLib Analysis Tools
Descriptive Statistics
Classification
Clustering
Visualization
Generate plots specific to analysis technique

3. EVALUATION

4. CONTINUING WORK

We identified the potential of acheiving better classification accuracy by training classifiers on subsets of data from regions that have similar distributions of defects. We propose adding a preprocessing stage to our platform that partitions the data by subdivision, computes the distribution of

© 2016 ACM. ISBN .
DOI:

defects for each subdivision, and then groups subdivisions based on a similarity metric for distributions.

Currently we use a simple similarity metric–the Chi-squared distance–and group subdivision based on a fixed similarity threshold. We plan to refine our similarity metric, use it to explore the efficacy of clustering techniques for grouping subdivisions, and evaluate the usefulness of the grouping by training and testing defect type classifiers on each such group.

Similarity metric: $S(D_1, D_2) = \sum_{i=0}^{N} \frac{(x_i - y_i)^2}{sdfdsfdsf}$

5. CONCLUSIONS

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] D. H. Schafer. A prediction model for broken rails and an analysis of their economic impact. *2008 AREMA Conference*, 2008.

[2] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 2–2, Berkeley, CA, USA, 2012. USENIX Association.

[3] A. M. Zarembski. Some examples of big data in railroad engineering. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 96–102, Oct 2014.