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

Dylan Chapp and Surya Kasturi
Advisors: Michela Taufer and Nii Attoh-Okine

Motivation

- Railroad network resiliency depends identification of defects
- Sensor-equipped monitoring cars collect rail and track-geometry data
- Can we use the data to predict defect occurrences in rail subdivisions?

Sensor-equipped Monitoring Car with its sensors

Rail and Track Defects Data Sets

- Rail defects data: physical degradation
  - 26,432 20-dimensional data points
- Track geometry data: misalignment
  - 25,421 41-dimensional data points
- Mixed numerical and categorical data

Scalable Reproducible Data Analysis Workflow

Track Defect Data

Rail Defect Data

Raw data provided by collaborators

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

Docker Container

Input Data

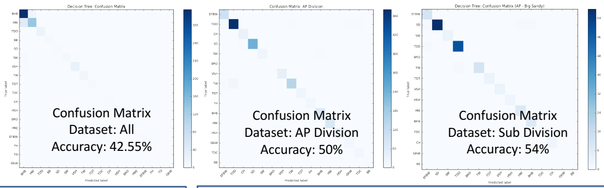
Analysis Code

- Package data, scripts, configs as a Docker container
- Collaborators can easily reproduce our analysis
- Hosted on DockerHub: https://hub.docker.com/r/dchapp/rail

Predicting Defect Types

Can we predict defects in railroad tracks?

- Defects are classified using Decision Tree
- Defect size, accumulated tonnage, rail weight, rail section age are used as features



Defect Super Classes

T – Class	TDD, TDC, TD, TDT
B – Class	BRO, BHB, BB
W – Class	HW, HWJ, OAW, SW

Defect Super Class T, B, W Classification

Defect TDD, TD, TDT Classification

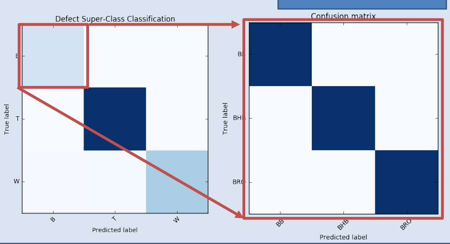
Defect BRO, BHB, BB Classification

Defect HW, HWJ, OAW, SW Classification

Pipeline

Defects are classified using super-class before predicting their type

Using the subset containing T, B, W defects resulted in accuracy of 89.90%

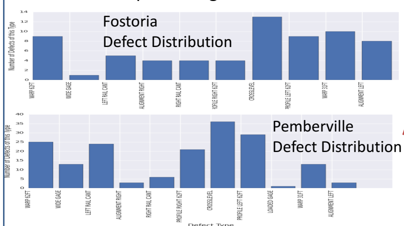


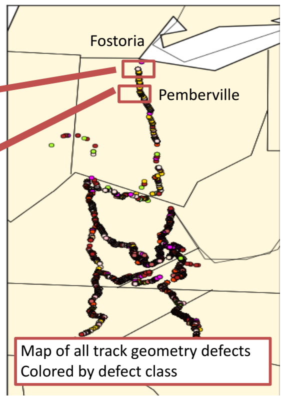
Conclusions:
- We improve prediction accuracy by a hierarchical classification scheme
- First decide membership in defect superclass, then in defect class using a second classifier

Track Region Similarity Analysis

Can track regions be grouped so that defect-type classifiers trained on region-specific data achieve better accuracy?

1: Extract each pair of regions' defect distributions

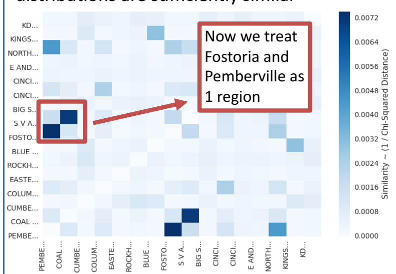




2: Compute the Chi-Squared Similarity of each pair of defect distributions

$$d(\mathbf{x}_u, \mathbf{x}_v) = \frac{1}{2}\sum_{n=1}^{N}\frac{[x_u(n) - x_v(n)]^2}{x_u(n) + x_v(n)}.$$

3: Group regions together whose defect distributions are sufficiently similar



Heatmap of regions' defect similarities

In progress:
- Training classifiers on subsets of defect data from statistically similar regions

References

1. A. Zarembski, "Some Examples of Big Data in Railroad Engineering", IEEE International Conference on Big Data, 2014
2. Track Inspector Rail Defect Reference Manual, Federal Railroad Administration, Rev. 2, 2015

Acknowledgements / funding