Discovering Bellwether for Defect Prediction using Hoeffding Races

Akshay Nalwaya (200159155)
Sanjana Kacholia (200263479)
Shantanu Sharma (200255931)

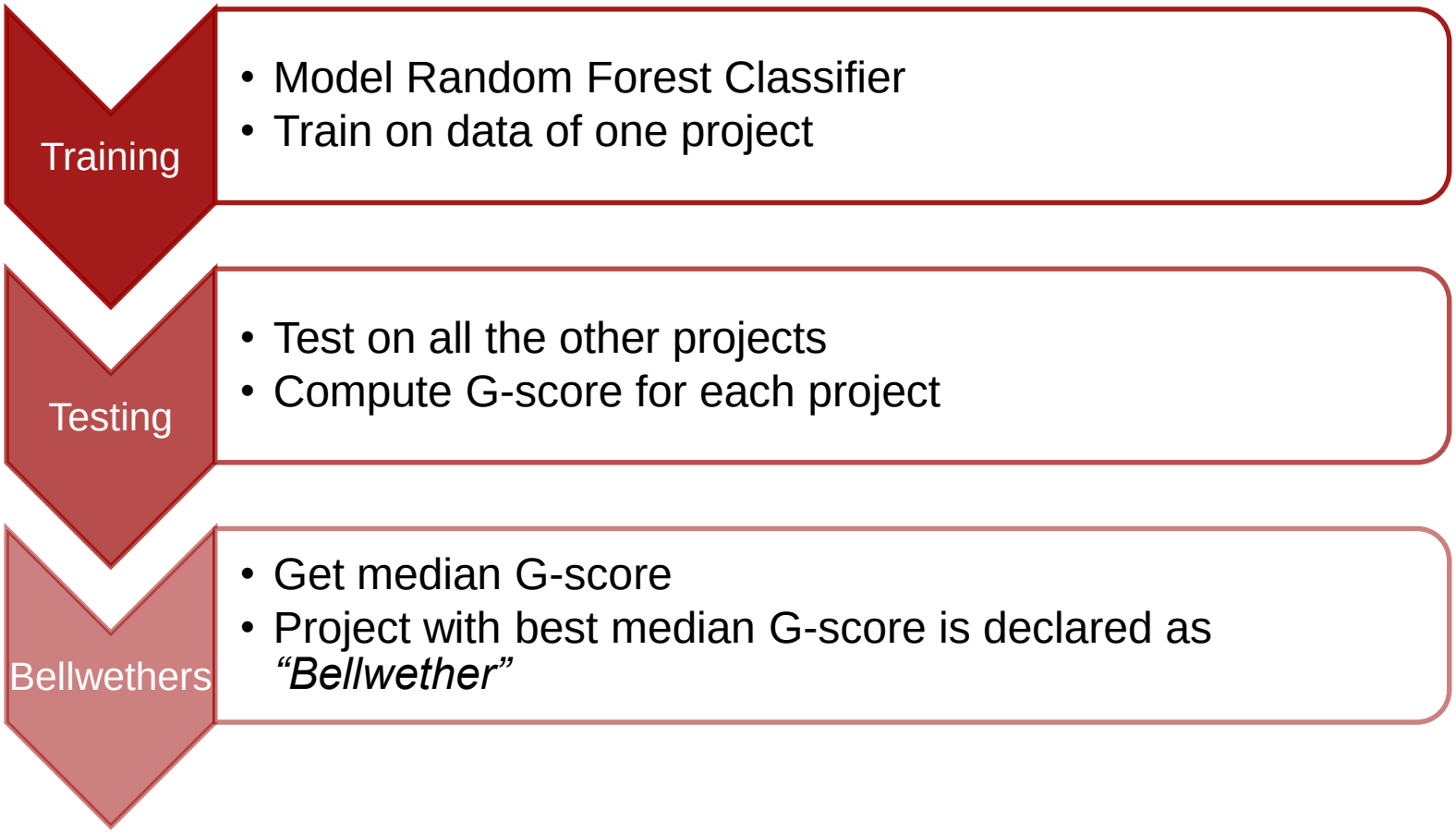
Overview

- *Why* : Identifying the Bellwether project among a group of projects
- *What* : Making the identification of this Bellwether project faster than the current $O(N^2)$ approach
- *How* : Use Hoeffding races to reduce the dataset required for training and/or testing for Bellwether identification

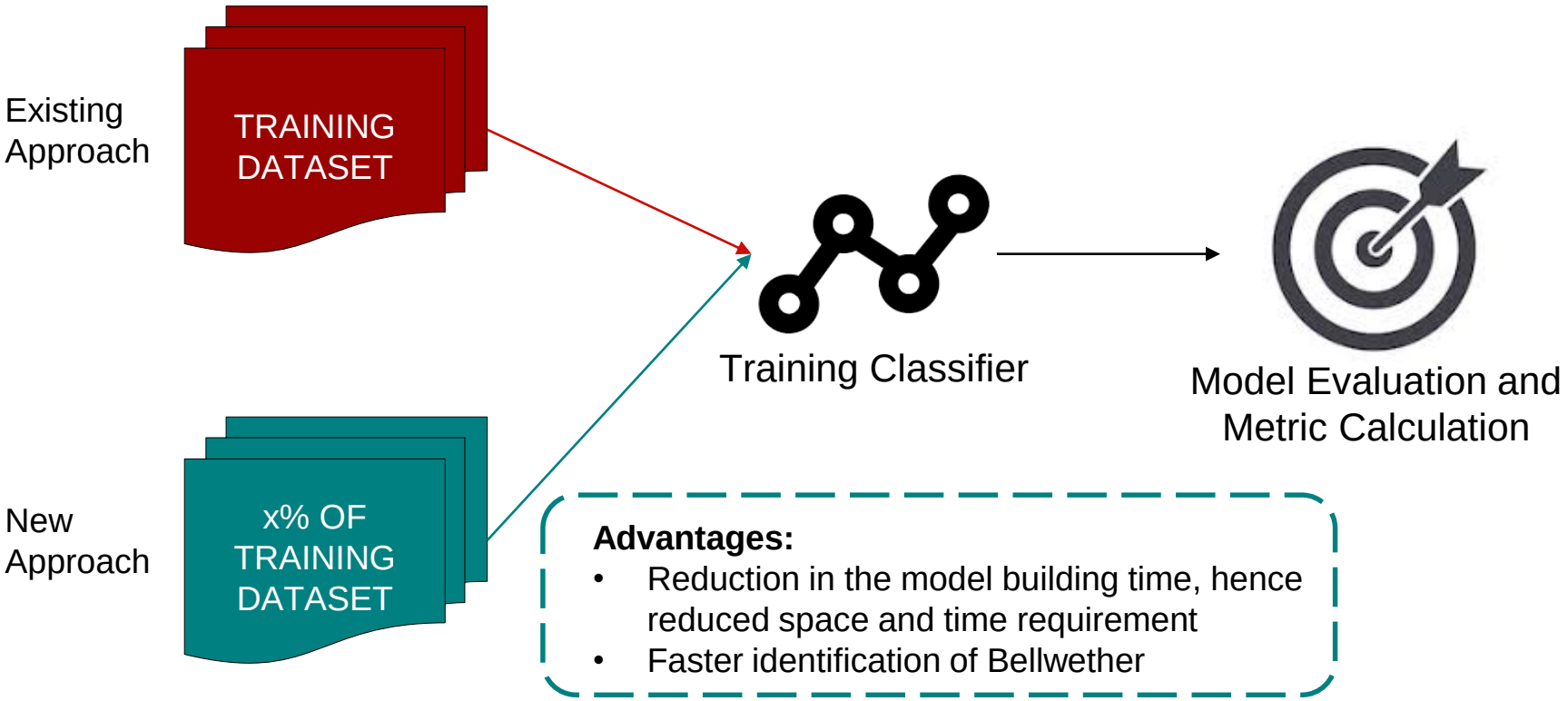
Why Bellwethers?

- Bellwethers just uses one dataset to construct prediction model
- Predicting defects for the new projects using data from the Bellwether project

Baseline Approach



Determining Hoeffding Bounds helps to reduce the training set



Research Questions

RQ1 : Can we predict which dataset will be the bellwether?

RQ2 : Can we reduce the time to find bellwether by reducing the size of data?

Results

Projects	Baseline G-score	New G-score
ant	0.16	0.16
camel	0.23	0.24
ivy	0.08	0.10
jedit	0.02	0.02
log4j	0.32	0.32
lucene	0.53	0.53
poi	0.59	0.59
velocity	0.49	0.47
xalan	0.56	0.57
xerces	0.42	0.42

*New approach resulted in increased G-score for **ivy** and **xalan** while the other remained nearly similar*

Projects	Training Data
ant	8.5%
camel	8.5%
ivy	8.5%
jedit	8.0%
log4j	8.5%
lucene	8.5%
poi	8.5%
velocity	8.0%
xalan	8.3%
xerces	8.0%

- ***'poi'*** is the bellwether dataset for the baseline method as well as after the implementation of Hoeffding races.
- Training data of around ~8.5% for each dataset gives similar results, **reducing the time and data required for training effectively**.

Other experiments for Bellwether identification

In addition to sampling just the training set we performed the following experiments for finding Bellwethers:

- Reducing only the testing dataset
- Reducing both the training and testing dataset

Future Work

- Using different racing algorithms and compare their performance with Hoeffding races
- Extending this work to different target domains like code smells, issue lifetime estimation and effort estimation

Questions?

References

- Bellwethers: A Baseline Method For Transfer Learning by Rahul Krishna et al (link)
- The Racing Algorithm: Model Selection for Lazy Learners by Maron et al (link)