CS6240: Final Project

Predict sightings of the Red-winged Blackbird in Birding Checklists

Utkarsh Jadhav
Sriharsha Srinivasa Karthik Kaipa

Table of Contents

- **Overview and Approach**

- Performance comparison
- Scope For Improvement

Overview and Approach

- Technologies used -
  - Spark
    - MLLib - Machine Learning Library
    - Scala - Functional Programming Approach
  - AWS EMR

- Approach for Classification
  - Random Forest Classification (Ensemble Method)
  - *Why* Ensemble?

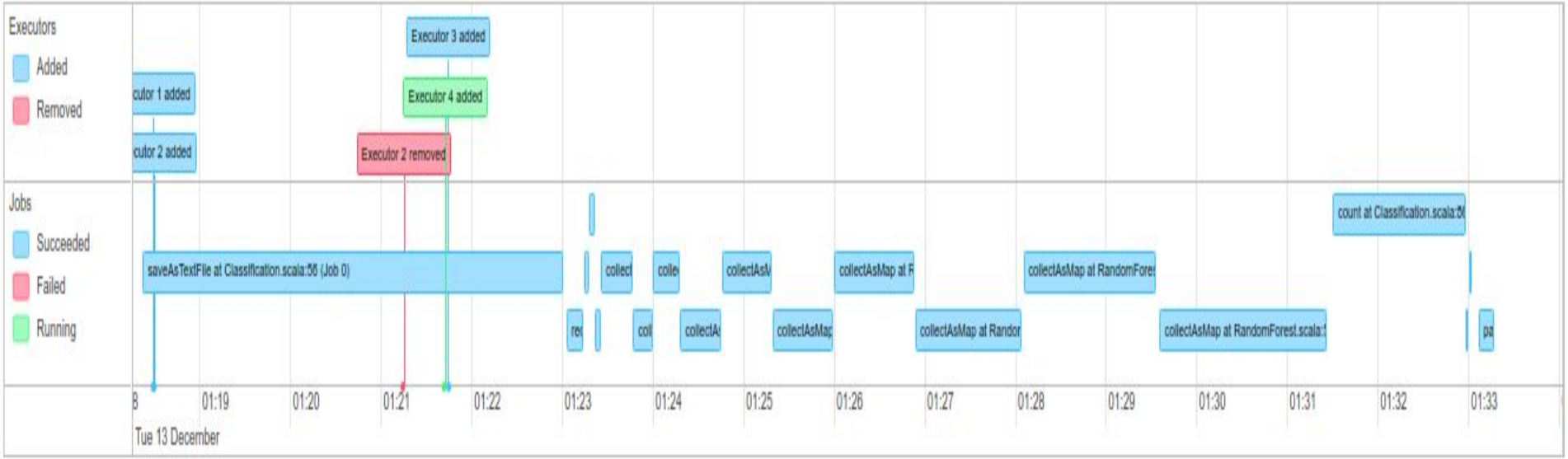
- Advantages of Spark
  - Easy to write , Scala
  - Concept - Partitioning , Repartitioning

Table of Contents

- Overview and Approach
- **Performance comparison**
- Scope For Improvement

Performance Comparision

- Total Execution Timeline



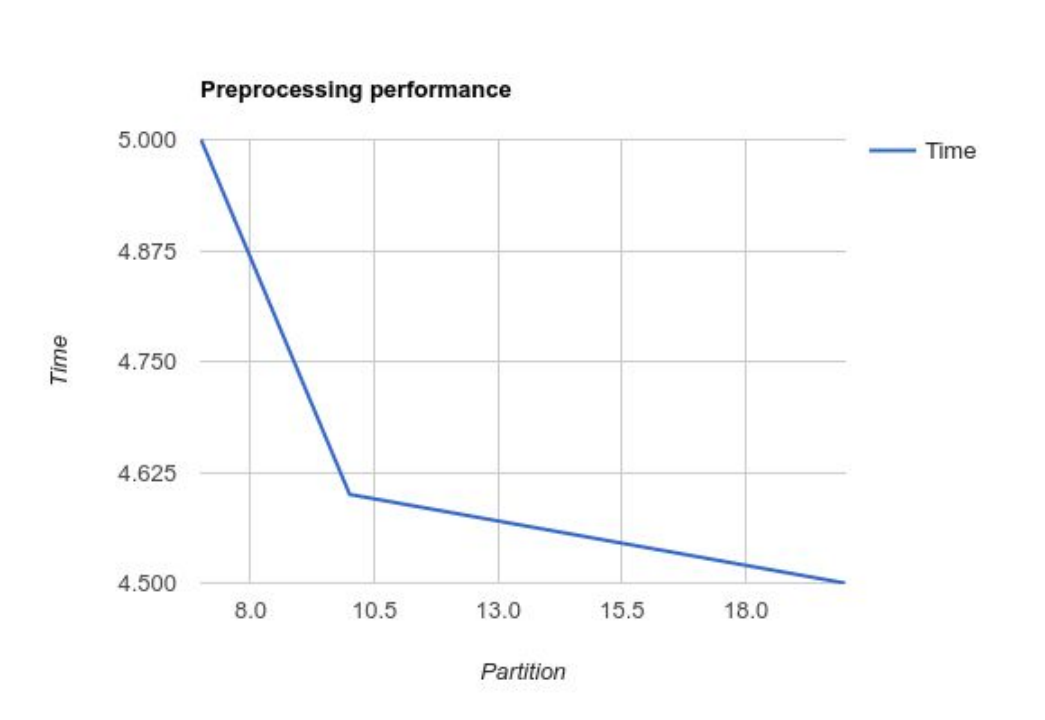
Performance Comparision

- Time per task

17	saveAsTextFile at treeEnsembleModels.scala:447	2016/12/13 01:33:00	0.7 s	1/1	
16	count at Classification.scala:86	2016/12/13 01:32:58	1 s	1/1 (1 skipped)	
15	count at Classification.scala:86	2016/12/13 01:31:30	1.5 min	1/1 (1 skipped)	
14	collectAsMap at RandomForest.scala:550	2016/12/13 01:29:35	1.8 min	2/2 (1 skipped)	
13	collectAsMap at RandomForest.scala:550	2016/12/13 01:28:05	1.5 min	2/2 (1 skipped)	
12	collectAsMap at RandomForest.scala:550	2016/12/13 01:26:53	1.2 min	2/2 (1 skipped)	
11	collectAsMap at RandomForest.scala:550	2016/12/13 01:25:59	53 s	2/2 (1 skipped)	
10	collectAsMap at RandomForest.scala:550	2016/12/13 01:25:18	40 s	2/2 (1 skipped)	
9	collectAsMap at RandomForest.scala:550	2016/12/13 01:24:45	33 s	2/2 (1 skipped)	
8	collectAsMap at RandomForest.scala:550	2016/12/13 01:24:18	27 s	2/2 (1 skipped)	
7	collectAsMap at RandomForest.scala:550	2016/12/13 01:23:59	18 s	2/2 (1 skipped)	
6	collectAsMap at RandomForest.scala:550	2016/12/13 01:23:46	13 s	2/2 (1 skipped)	
5	collectAsMap at RandomForest.scala:550	2016/12/13 01:23:25	22 s	2/2 (1 skipped)	
4	collectAsMap at RandomForest.scala:894	2016/12/13 01:23:21	3 s	2/2 (1 skipped)	
3	count at DecisionTreeMetadata.scala:116	2016/12/13 01:23:17	4 s	1/1 (1 skipped)	
2	take at DecisionTreeMetadata.scala:112	2016/12/13 01:23:14	4 s	2/2	
1	reduce at MLUtils.scala:92	2016/12/13 01:23:02	11 s	1/1	
0	saveAsTextFile at Classification.scala:56	2016/12/13 01:18:22	4.6 min	2/2	

Performance Comparision

- Preprocessing performance scale-up



Performance Comparision

- Model training + testing performance scale-up

Performance Comparision

- Total performance scale-up

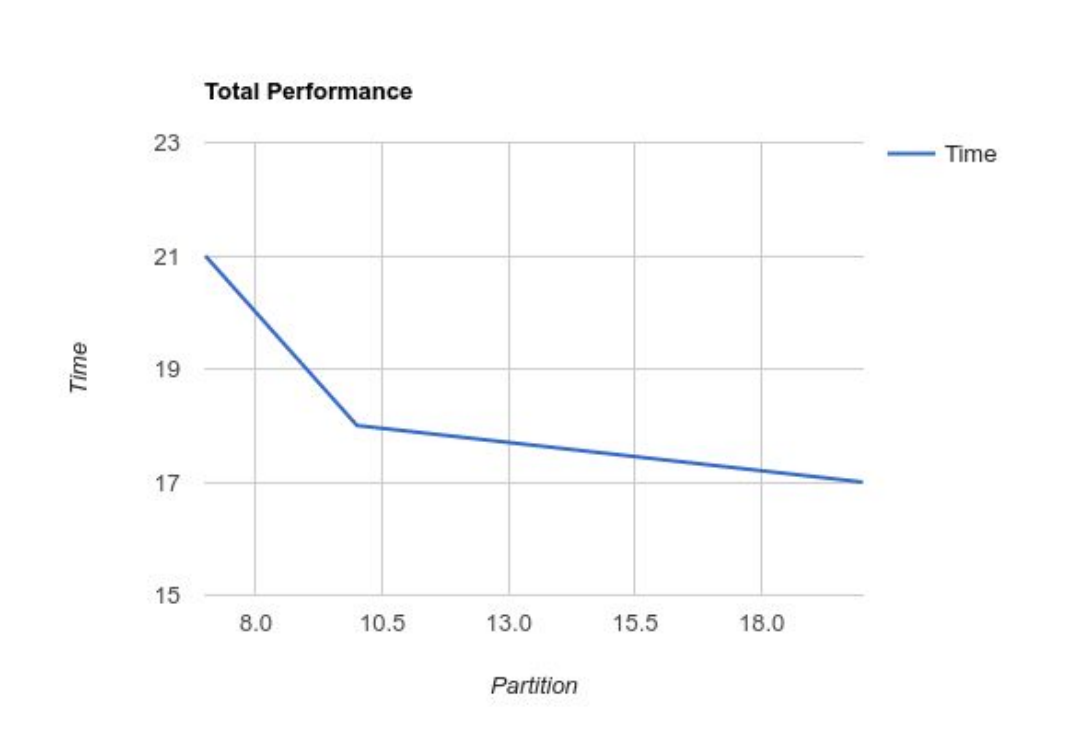


Table of Contents

- Overview and Approach
- Performance comparison
- **Scope For Improvement**

Scope For Improvement

- Emphasis on Data Mining Techniques
  - Attribute Ranking
  - Removal of bias, etc

- MLLib is black-box! Generalization is harmful!

Thank you!

Questions?