BIG DATA RANDOM FORESTS IN R

Andrew Nisbet

April 2016

MSA220 project 1

PROJECT

Aim: investigate using R to perform random forest classification on a large dataset.

- Predict delays from flight information
- Compare `randomForest` and `bigrf`
- Evaluate performance

OVERVIEW

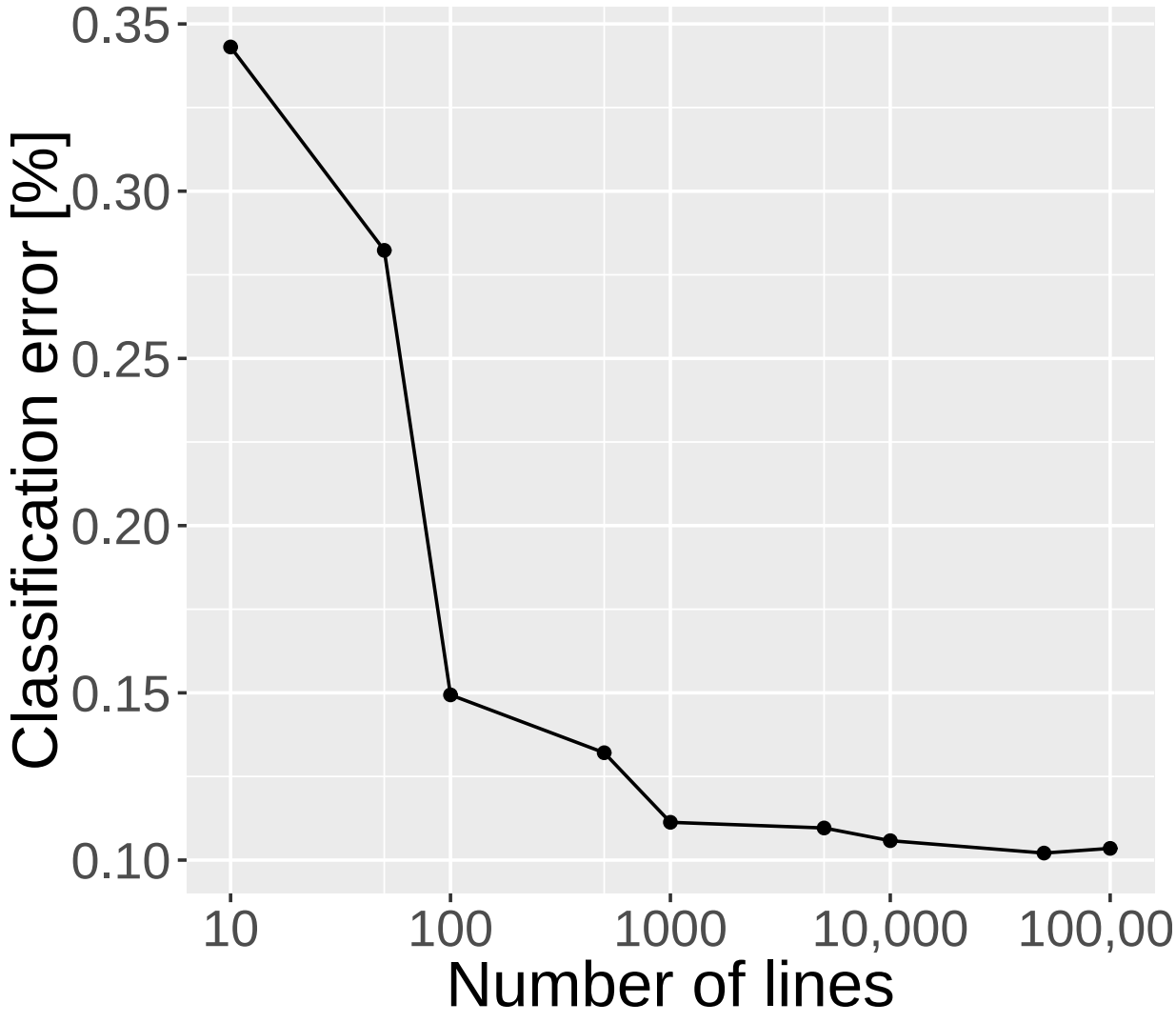
- About big data
- Dataset creation
- `randomForest`
- `bigrf`
- Conclusion

BIG DATA - RAM

Keep everything in RAM, otherwise use disk intelligently.

	Hard Drive	RAM
Transfer Rate	$50\,\mathrm{MB\,s^{-1}}$	$5000\,\mathrm{MB\,s^{-1}}$
Access Delay	10 000 000 ns	10 ns

BIG DATA - MOTIVATION



DATASET - OVERVIEW

- Combination of flight (12 GB) and weather (3 GB) data.
- Fight features

  `Day of the week, duration, airport`
- Weather features
- Only flights leaving a single airport
- y = whether flight was delayed
- Final dataset had 5 000 000 lines, 19 features, 500 MB
- big-n

DATASET - SUBSETTING

Large datasets distrup standard workflows. Methods to create a 5 000 000 line subset:

- `read.table`: crashed
- `read.table`: 250 s
  - Read column types in advance
  - Specify `nrows` (from `wc -l input.csv`)
- `fread`: 20 s (`data.table`)
- `shuf`: 2 s (Linux command line)

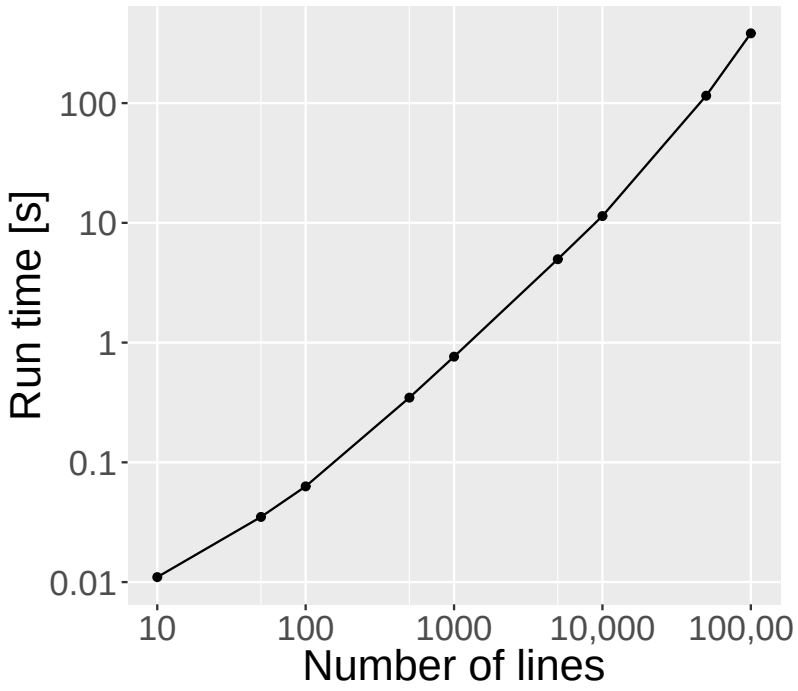
```
shuf -n 10000 input.csv > output.csv
```

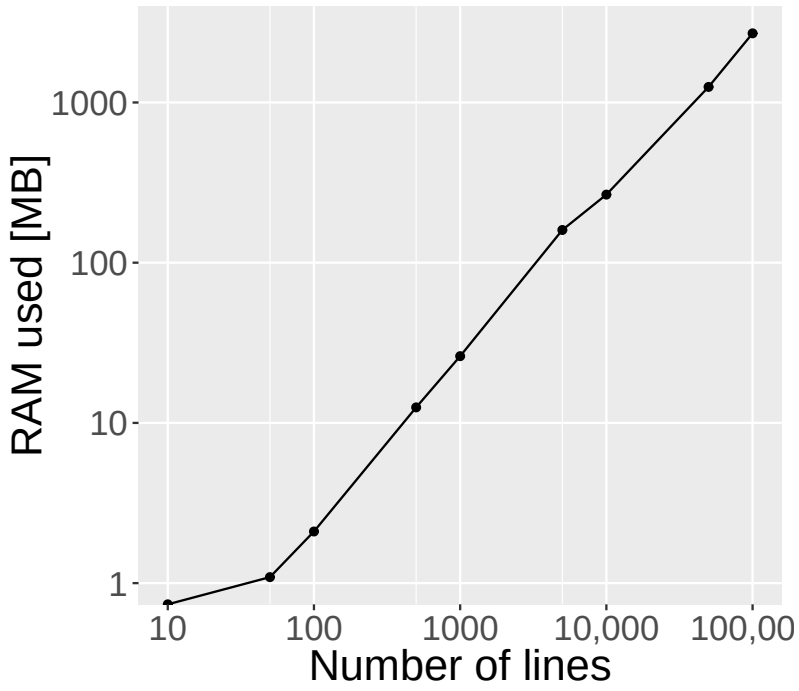
randomforest - OVERVIEW

Most common package

- Limited to 53 factor levels
- Single process

randomforest - RESULTS



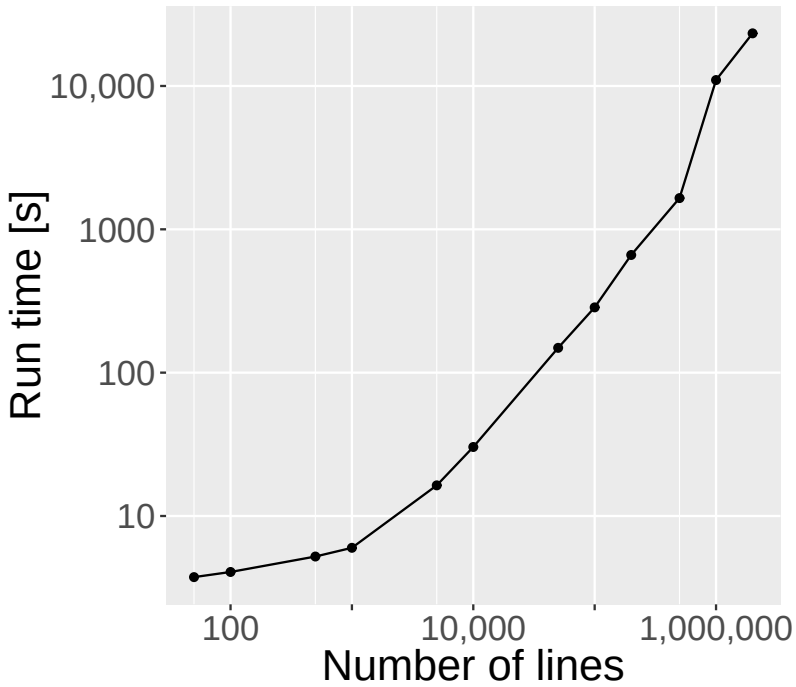


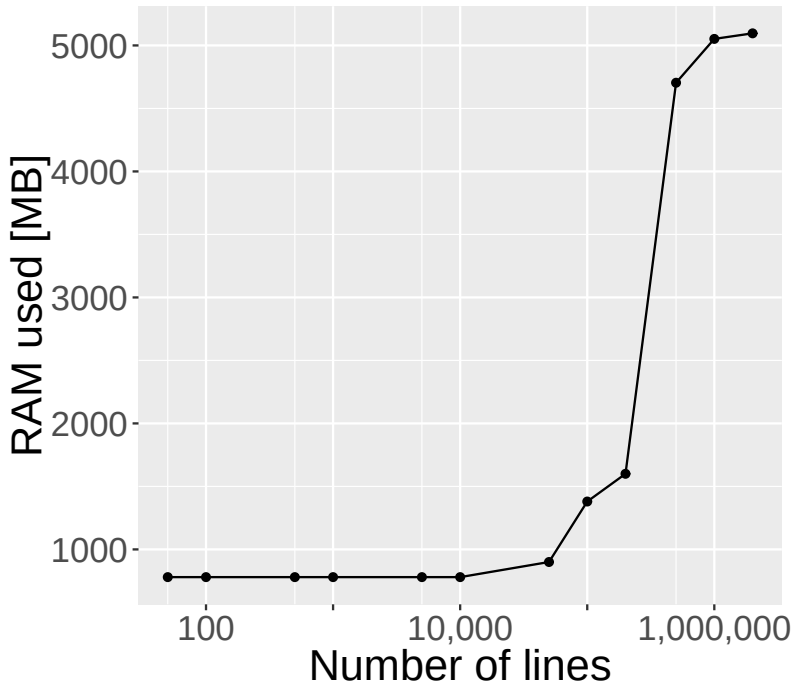
Crashed after 100 000 lines.

bigrf - OVERVIEW

- Uses `big.matrix` for data and trees
- Data shared between processes
- Data can be saved to disk and queried
- Outputs useful statistics
- Linux/OSX only
- Lack of support and documentation

bigrf - RESULTS





Crashed after 2 000 000 lines.

CONCLUSION

- R can be used for large datasets, just choose the right tools
- `bigrf` is nice to work with, compared to `randomForest`
- `bigrf` can handle much larger datasets without crashing, but performs poorly for small datasets

Questions?