Detecting Bots on an Online Auction

Alexander Chapman Kowaliw, Andrej Rupinski, Bradley Gallagher, Daniel Kenny, Maciej Sielski, Roni Habib

Abstract

- What is the problem of the dataset?
- Why is a big data solution needed?

Introduction/Background:

- Where did the data come from & what would its solution be used for?
- Why is this a big data problem?
- What type of problem is this (e.g. classification/regression)?
- Literature review – past solutions and problems

Methodology:

- Local vs global approaches
- ML Model comparison on pipeline performance
- Big data pipeline outline (image & explanation)
- Evaluation metrics used: model accuracy, and pipeline performance (computational cost/requirements - potential real-world cost if many partitions are optimal)

Experiments:

- Testing different dataset sizes and partition quantities to extrapolate correlation between computational performance and scale out/speed up.
- Testing different approaches (global/local) to determine the effect on model/pipeline performance.

Results:

- What models and pipelines performed the best?
- What made them perform the best?
- How do they compare to each other (comparison of extrapolations as well)?
- Could this pipeline be applied to different data sets of similar problems? (Can this be included in experiments?)

Conclusion:

- What was found with model and pipeline approach testing? (computational performance & extrapolation for scale out/speed up)
- Did one pipeline permutation perform significantly better or worse than others? If so, why?

References: