The Logic of Benchmarking: A Case Against State-of-the-Art Performance

Wheeler Ruml

University of New Hampshire

(thanks to the NSF RI and the DARPA CSSG programs for support)

Three Points

We should prefer

1. to solve small benchmarks than large ones
2. to understand performance than to achieve state-of-the-art
3. to perform reasonably in many domains than to excel in one

Actual Comments

"I definitely agree that toy problems are relevant. In fact, that's the only thing I do: the N-puzzle, Pancake, Rubik's cube and the like :) but I never solve the 8-puzzle but work on the 24-puzzle instead, hope you see my point. Solving problems in the scale of microseconds does not seem very useful to me"

— reviewer

"I found it unfortunate to allow the search algorithms to run for only 5 minutes. If one is expected to derive conclusions on the general trend of a search algorithm I would encourage you to run them for longer"

— reviewer

Smaller Benchmarks are Better

Advantages:

- faster to run
- easier for others to reproduce
- allows wider variety of instances
- allows more detailed understanding of performance

Perceived disadvantages:

- measurement error
- not ‘real’ enough
- scaling not evident
- different phenomena at large scale

Actual Comments

"I believe that any publishable paper should demonstrate at least one domain on which the authors' algorithm outperforms the previous state of the art."

— reviewer

The Purpose Is Understanding

The goal is not to solve our toy problems.

The goal is to predict behavior on new (complex) problems.

Developing this predictive understanding is implicitly discouraged by emphasizing state-of-the-art performance.

The Purpose Is Understanding

The goal is not to solve our toy problems.

The goal is to predict behavior on new (complex) problems.

Developing this predictive understanding is implicitly discouraged by emphasizing state-of-the-art performance.

Should we require results that show a technique failing?

Actual Comments

"Isn't heuristic search just a topic for textbooks?"

— AAAI Fellow

"I am concerned that readers of [journal] will get the impression that heuristic search has become a community that is only interested in peculiarities of the 15-puzzle, talks only to itself, and has no relevance to broader AI or CS research."

— reviewer

General-purpose Techniques Are Better

Different goals:

1. best performance on problem X
2. reasonable performance on any problem like X
   - state-of-the-art search spaces are unnecessary, misleading
   - special requirements are disadvantages

GAs, annealing, and local search are very popular — why?

Where are our industrial sponsors?

Recap

We should prefer

1. to solve small benchmarks than large ones
2. to understand performance than to achieve state-of-the-art
3. to perform reasonably in many domains than to excel in one