Statistical Methodology for Comparison of SAT Solvers

Mladen Nikolić
Automated Reasoning GrOup
University of Belgrade

SAT 2010, UK

July 13, 2010.

Overview

1. Introduction
2. Statistical hypothesis testing
3. Methodology
4. Example evaluation
5. Related work
6. Conclusions

Overview

1 Introduction

2 Statistical hypothesis testing

3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

Comparison of SAT solvers

- Importance of SAT solver comparison
  - Significant number of proposed modifications each year
  - Their usefulness is not always self-evident
  - Need to discriminate better between good and bad ideas
- The approach most often used
  - Can be unreliable
  - Can't decide if the observed difference could arise by chance
  - Doesn't use solving times to the full extent

Solver runtime variation

- Solving time of a solver on a formula can vary
- Each formula should be solved several times in order to sample from the runtime distribution
- What is a reasonable way of sampling?
  - Shuffling
  - Changing the random seed
  - Maybe even introducing very small changes to solver parameters?

Number of solved formulae can vary

- Solvers from Minisat hack track 2009
- Industrial instances (2009), graph coloring instances (2002)
- Cutoff time of 1200s
- 50 runs per formula

	Industrial		Graph coloring	
Solver	Max	Min	Max	Min
MiniSAT 09z	161	111	180	157
minisat_cumr r	156	107	190	180
minisat2	141	93	200	183
MiniSat2hack	144	93	200	183

- Variation of the number of solved formulae can be large

Variation in solver comparison

- For each pair of solvers, 10000 simulated comparisons were made on each benchmark set with shuffled variants chosen on random
- MiniSAT 09z vs. minisat_cumr r on industrial **92%:8%**
- minisat2 vs. MiniSat2Hack on industrial **6%:94%**
- minisat2 vs. MiniSat2Hack on graph coloring **74%:26%**
- The results of the comparison may vary due to solver runtime variation

Main goal

Make steps towards:

- Eliminating chance effects from the comparison
- Giving an information on statistical significance of the difference
- Making a better use of the solving data

Main difficulties

- Censored observations (cutoff time is given)
- Comparison of runtime distributions for each instance is required
- Combining conclusions obtained on individual instances

Overview

1 Introduction

2 Statistical hypothesis testing

3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

Statistical hypothesis testing

- Null hypothesis H_0 (e.g. no difference in solver performance)
- Test statistic T (e.g. some measure of difference in solver performance)
- p-value — the probability of obraining observed or more extreme value of T if H_0 were true
- If $p < \alpha$ then reject H_0
- Effect size (sometimes the value of T will do)

Overview

1 Introduction

2 Statistical hypothesis testing

3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

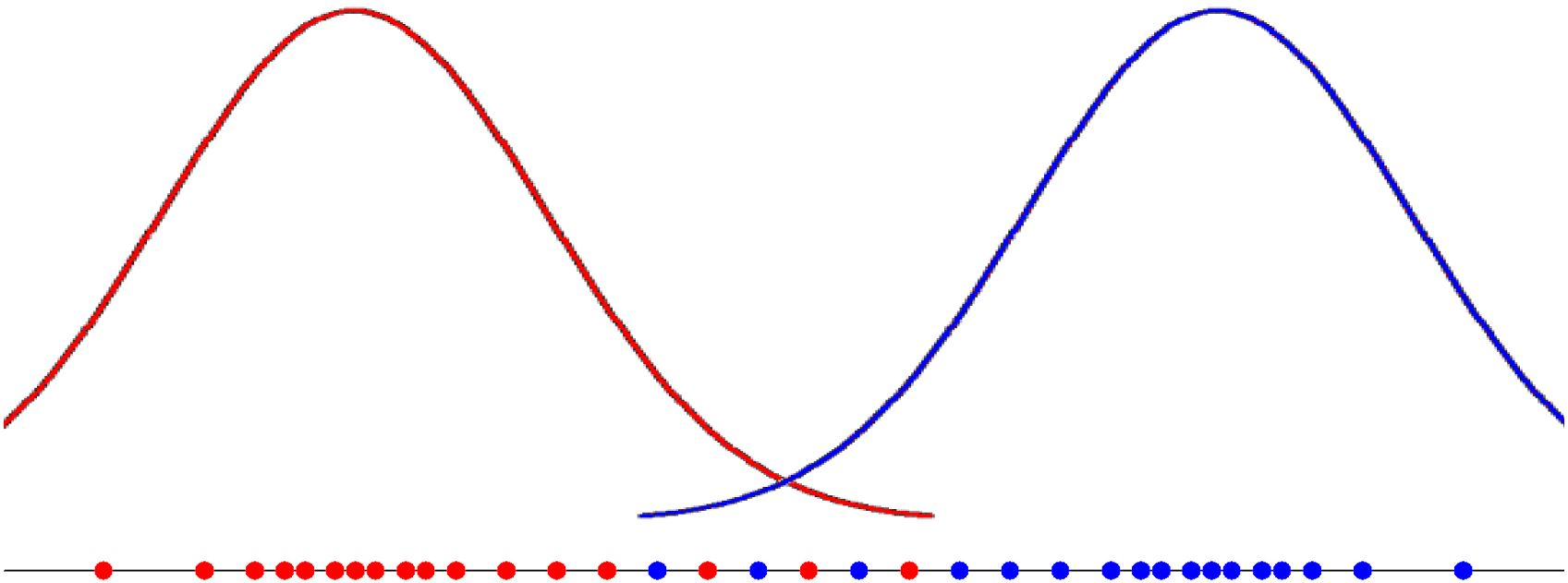
Overview of the methodology

Comparison of two solvers:

- Null hypothesis H_0: no difference in solver performance
- For each formula F_i take samples of runtimes A_i and B_i for each solver
- Calculate difference $d(A_i, B_i)$ for all i
- Calculate the average $\overline{d}$ of d values (it shouldn't deviate too much from its expectation under the null hypothesis $E_{H_0}\overline{d}$)
- Estimate the p value (by measuring the probability of the deviation)
- If $p < \alpha$ we judge which solver is better by the sign of $\overline{d} - E_{H_0}\overline{d}$

Choice of function d

- What could be a good choice for the function d?



- Estimate of $P(X < Y)$ suitable for censored data?

Determining statistical significance and the effect size

- How is the average $\overline{d}$ distributed?
- The distribution of $\overline{d}$ is asymptotically normal with parameters that can be estimated from the data
- For effect size measure we take $\overline{d}$ — the expected (over the formulae of the corpus) probability of one solver being faster than the other

Ranking

- If there is more than 2 solvers, ranking can be produced from pairwise comparisons
- Kendall-Wei method

Overview

1 Introduction

2 Statistical hypothesis testing

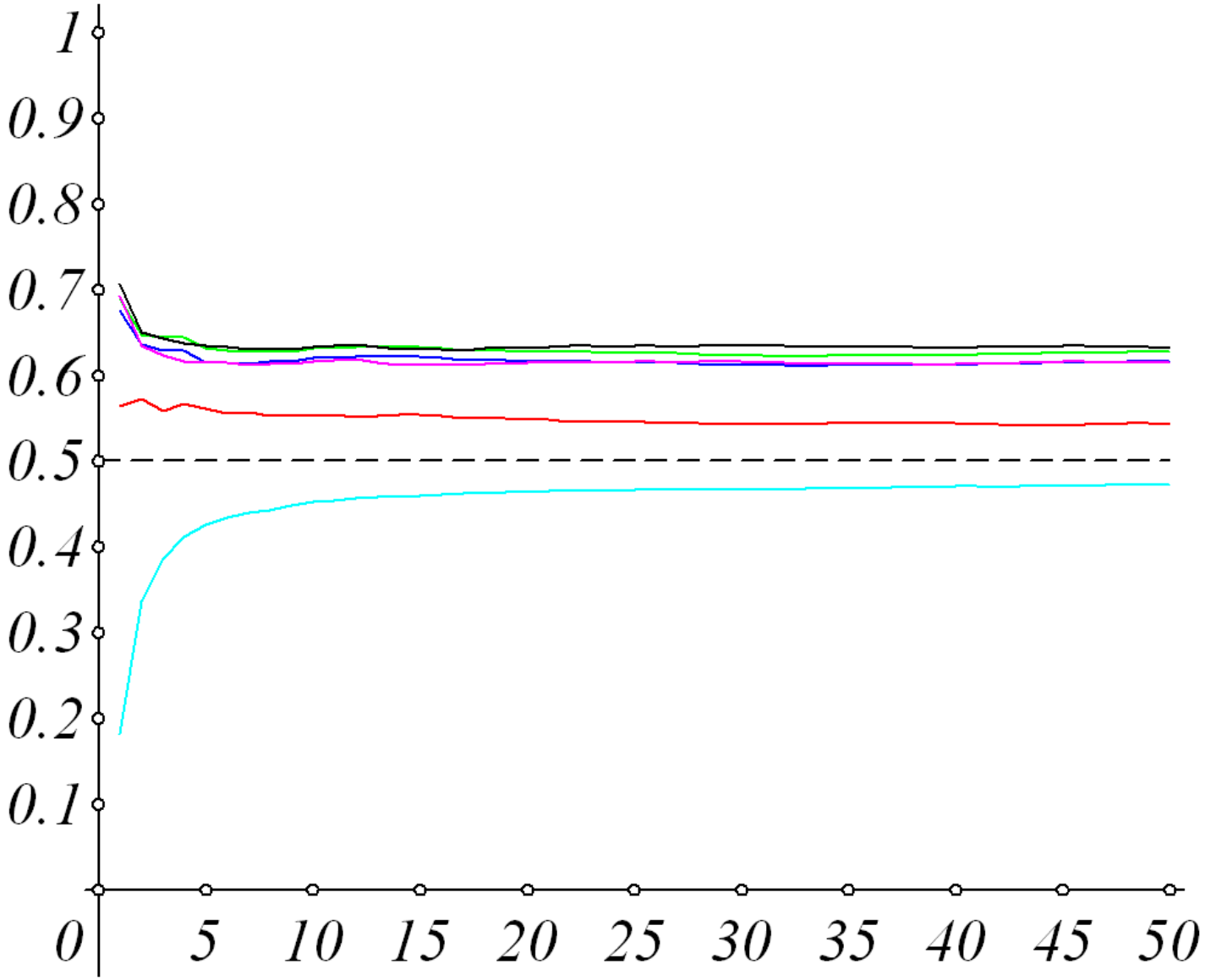
3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

How many runs per formula do we need?



Results of comparison on industrial instances

- $\alpha = 0.05$
- S_1 — MiniSAT 09z
- S_2 — minisat_cumr r
- S_3 — minisat2
- S_4 — MiniSat2hack
- All the differences on industrial instances are statistically significant, on graph coloring, some are not ($\alpha = 0.05$)

$P(X < Y) - 0.5$	S_1	S_2	S_3	S_4
S_1	-	0.055	0.134	0.123
S_2	-0.055	-	0.131	0.113
S_3	-0.134	-0.131	-	-0.040
S_4	-0.123	-0.113	0.040	-

Overview

1 Introduction

2 Statistical hypothesis testing

3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

Related work

- Le Berre, Simon (2003) — shuffling might be important for SAT solver comparison
- Audemard, Simon (2008) — shuffling can cause a large variation of the number of solved formulae
- Franc Brglez, et al. (2005, 2007) — use of standard statistical tests to compare two solvers on one formula and determine statistical significance
- ...

Overview

1 Introduction

2 Statistical hypothesis testing

3 Methodology

4 Example evaluation

5 Related work

6 Conclusions

Conclusions

- Advantages
  - Offers more reliable, statistical, information
  - Makes better use of the solving times
  - Could make identifying good ideas easier
- Drawbacks
  - The method is more complex and harder to understand
  - Higher computational cost (could be acceptable)
  - Doesn't use solving times to the full extent
- Open question
  - What is the most reasonable way to sample from the solver runtime distribution?

Thank you!