Methodology for Comparison and Ranking of SAT Solvers

Mladen Nikolić

Third Workshop on Formal and Automated Theorem Prooving and Applications

January 29, 2010.

Overview

Overview

Comparison of SAT solvers

- SAT solvers
- Importance of SAT solver comparison
  - Large number of proposed modifications each year
  - Their usefulness is not self-evident
  - We need to discriminate better between good and bad ideas
- Current approach
  - Unreliable
  - Sometimes inconclusive
  - No discussion if the observed difference could arise by chance

Motivation

	Graph coloring		Industrial	
Solver	Best	Worst	Best	Worst
MiniSAT 09z	180	157	159	112
minisat_cumr r	190	180	150	108
minisat2	200	183	140	93
MiniSat2hack	200	183	141	94

Main goals

- Eliminate chance effects from the comparison
- Decide if there is an overall positive or negative effect
- Give an information on statistical significance of the difference

Main difficulties

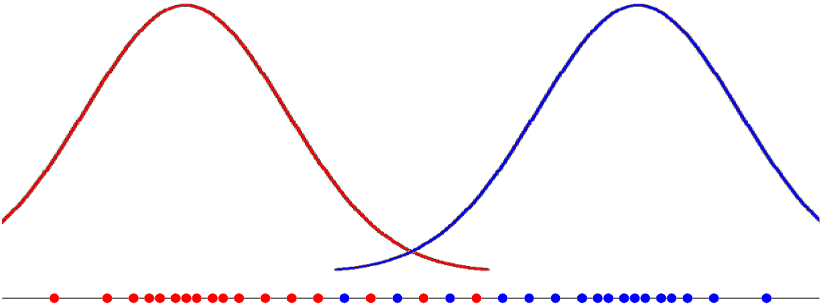
- Censored observations
- Comparison of distributions of solving times for one instance
- Combining conclusions obtained on individual instances

Overview

Statistical hypothesis testing

- Null hypothesis H_0
- Test statistic T
- $p = P(|T| \geq t | H_0)$
- If $p < \alpha$ then reject H_0
- Effect size

Comparing two distributions



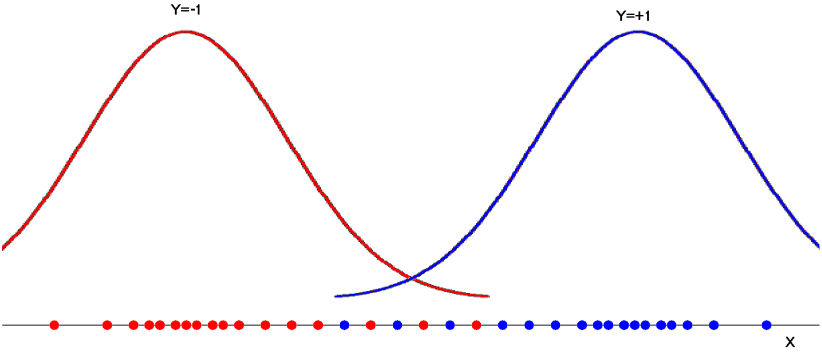
Point biserial correlation

- Point biserial correlation ρ_{pb} can be estimated by

$$r_{pb} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}$$

- $\rho_{pb}, r_{pb} \in [-1, +1]$

Point biserial correlation



Handling censored data

- Gehan statistic W_G
- $E(W_G) = P(X > Y) - P(X < Y)$
- $\frac{1-E(W_G)}{2} = P(X < Y)$

Overview

Sketch of the methodology

- H_0: no difference in solver performance
- Choose the level of statistical significance α
- Calculate differences d_i between samples of solving times of F_i
- Under the null hypothesis the average of d_i shouldn't be too large
- Estimate the p value and check the significance of the average difference
- Check and interpret the effect size

Choice of function d

- What could be a good choice for function d?
  - ρ_{pb}?
  - $\pi = P(X < Y)$?

Choice of function d

Theorem

Under some reasonable conditions the following relations hold

$$W_G = \frac{S_R S_Y}{n_1 n_2} r_{pb} \tag{1}$$

$$\frac{var(W_G)}{\frac{S_R^2 S_Y^2}{n_1^2 n_2^2} var(r_{pb})} \to 1 \quad (n_1 + n_2 \to \infty) \tag{2}$$

where

$$S_X = \sqrt{\sum_{i=1}^{n_1+n_2} (X_i - \overline{X})^2}$$

Determining statistical significance

- How is the average of d_i distributed (choosing r_{pb} for d_i)?

$$\bar{z} = \frac{1}{M} \sum_{i=1}^{M} z(r_i)$$

$$\bar{z} \sim \mathcal{N}\left(\frac{1}{M} \sum_{i=0}^{M} z(\rho_i), \frac{1}{M^2} \sum_{i=1}^{M} \frac{var(r_i)}{(1 - r_i^2)^2} \right)$$

Determining effect size

- Averages of estimates of ρ_{pb} or π on individual formulae

Ranking

- Potential problems with transitivity
- $P(A > B) > \frac{1}{2}, P(B > C) > \frac{1}{2} \Rightarrow P(A > C) > \frac{1}{2}$
- Kendall-Wei method

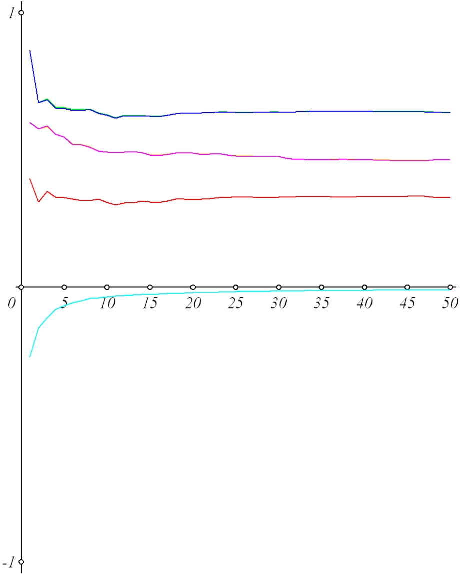
Overview

Results of comparison

- $\alpha = 0.05$
- Only the difference between S_3 and S_4 is insignificant

	ρ_{pb}				π			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
S_1	-	0.326	0.636	0.636	-	0.320	0.140	0.141
S_2	-0.326	-	0.465	0.464	0.680	-	0.239	0.239
S_3	-0.636	-0.465	-	0.010	0.860	0.761	-	0.506
S_4	-0.636	-0.464	-0.010	-	0.859	0.761	0.494	-

How many shuffles do we need?



Overview

Related work

- Daniel Le Berre, Laurent Simon (2004) — shuffling might be important for SAT solver comparison
- Franc Brglez, et al. (2005, 2007) — use of standard statistical tests to compare two solvers on one instance yielding p value (statistical significance)

Overview

Conclusions

- Current approach is unreliable
- New, statistically founded, methodology
  - Offers more reliable information
  - Could make identifying good ideas easier
- Total computational cost can actually stay the same