Statistical Methodology for Comparison of SAT Solvers

Mladen Nikolić*
Faculty of Mathematics, University of Belgrade, Serbia

Abstract

Evaluating improvements to modern SAT solvers and comparison of two arbitrary solvers is a challenging and important task. Relative performance of two solvers is usually assessed by running them on a set of SAT instances and comparing the number of solved instances and their running time in a straightforward manner. In this paper we point to shortcomings of this approach and advocate more reliable, statistically founded methodologies that could discriminate better between good and bad ideas. We present one such methodology and illustrate its application.

1 Introduction

Many SAT solvers have been developed and various improvements to them have been proposed over the years, especially in the domain of heuristic components.

In order to assess the quality of a proposed modification, one usually runs a modified and the base version of the solver on some set of SAT instances. The solver that solves more instances, or the same number of instances in less time is considered to be better. This approach can be flawed because solving times of instances can significantly vary depending only on trivial properties of the formula like ordering of clauses and literals, or on random seeds used, which can lead to different experimental results by chance.

We performed experiments to investigate this claim using four solvers were chosen from the MiniSAT hack track of the SAT 2009 competition and two benchmark sets — the first consisting of 292 industrial instances used at the MiniSAT hack track and the second consisting of 300 graph coloring instances from the SAT 2002 competition. Each solver was run on 50 shuffled variants of each benchmark (obtained by reordering the clauses, literals in each clause, and renaming the variables) with cutoff time of 1200 seconds. Each two solvers were compared on both benchmark sets. In three comparisons the probabilities of solvers swapping places when the shuffled variants of formulae were chosen on random weren't negligible (6%-26%). Also, observed variation of number of solved formulae was large. More information can be found in [Nik10].

In addition to the problem just discussed, there is a problem of drawing conclusions from the available experimental results. Sometimes, the results are presented by tables showing that the new SAT solver is performing better than the base one on some subsets of instances, and worse on the others, without clear conclusion about the overall effect. Also, SAT solver comparisons are concluded without discussion if the observed differences could be obtained by chance or are a consequence of a genuine effect.

The goal of this work is the formulation of statistically founded methodology of SAT solver comparison that would *i*) eliminate chance effects from the results, *ii*) give an answer if there is a positive (or negative) overall effect of the proposed modification to SAT solver performance, and *iii*) give an information of statistical significance of that effect. Such a methodology would

*This work was partially supported by Serbian Ministry of Science grant 144030.

enable more reliable discrimination between good and bad ideas, enabling the community to focus on the more promising ones.

There are several issues that have to be addressed in devising such methodology. The first is a presence of censored data. If the formula is not solved in a given cutoff time, it is only known that it needs more time to be solved, but not how much exactly. The second is a need to compare runtime distributions instead of single solving times that are unreliable. The third issue is finding a way to combine conclusions for different formulae to derive an overall conclusion.

The rest of the paper is organized as follows. The proposed methodology is sketched in Sect. 2 and the experimental results are given in Sect. 3. In Sect. 4, related work is discussed. In Sect. 5 final conclusions are drawn. A more detailed description of the methodology and the concepts used can be found in [Nik10].

2 The Methodology

Let random variable τ^j represent runtimes of the solver S_j $(j = 1, 2)$ on SAT instance F. Since solving times can be too large for practical evaluation, a cutoff time T is used, and thus distributions of random variables τ^j are truncated to the right at the point T. The difference of SAT solver performances should be defined by some function $\delta(\tau^1, \tau^2)$ measuring the suitably chosen difference between distributions of these variables. Since the random variables themselves are not available, inferences about them are made using samples of runtimes. The value of the function δ should be approximated by a difference d between samples. The differences δ_i of random variables corresponding to formulae F_i can be averaged to obtain a value $\overline{\delta}$ which measures the overall difference between solvers on given corpus of formulae. Sample estimate of $\overline{\delta}$, the average of d_i values, will be denoted $\overline{d}$. Distribution of the average of $\overline{d}$ under the hypothesis $\overline{\delta} = 0$ will be denoted by Θ.

The methodology is outlined in Fig. 1. It can be considered as a statistical test with the null hypothesis that there is no overall effect — H_0: $\overline{\delta} = 0$.

Obviously, in order to use this methodology, its various aspects must be discussed. The most important ones are the choice of the function d, estimation of distribution Θ, and interpretation of the magnitude of $\overline{d}$. We will propose some choices for each of these aspects.

The role of function d is to quantify the difference in performance of two solvers on one instance based on samples of corresponding solving times. For that we use effect size measures for difference between two samples.

As an indicator that two solvers perform the same on some instance F, we take

$$P(\tau^1 > \tau^2) = P(\tau^1 < \tau^2)$$

or equivalently

$$\omega = P(\tau^1 > \tau^2) - P(\tau^1 < \tau^2) = 0$$

where τ^j is a random variable representing solving times of the solver S_j on instance F. These two probabilities need not sum to 1 in case that censored data are present. In that case

$$\pi = \frac{1 - \omega}{2} = P(\tau^1 < \tau^2) + \frac{1}{2} P(\tau^1 = \tau^2)$$

which is a quite intuitive measure that combines the evidence of one solver performing better than the other with the uncertainty that appears if both solvers haven't solved the same benchmarks. ω can be estimated by Gehan statistic W_G [Geh65, Man67].

- INPUT: Solvers S_1 and S_2, and the set of benchmark instances
- OUTPUT: Information if one solver is better than the other and estimate of the effect size
- Choose the level of statistical significance α ($\alpha < 1$)
- For each formula F_i from corpus $\mathcal{F}$ consisting of M instances:
  - Take a sample T_i^j of size N of random variable τ_i^j ($j = 1, 2$)
  - Calculate the difference $d_i = d(T_i^1, T_i^2)$ between obtained solving times
- Calculate the average $\overline{d}$ of values d_i
- Estimate Θ — the distribution of $\overline{d}$ under the null hypothesis
- Calculate the p value for $\overline{d}$ according to the distribution Θ
- If $p \leq \alpha$
  - Declare the first solver to be better if $\overline{d} < 0$
  - Declare the second solver to be better if $\overline{d} > 0$
  - Report $\overline{d}$ as the estimate of the magnitude of the difference between performances of two solvers
- otherwise, declare the difference insignificant

Figure 1: Outline of the proposed methodology.

Point biserial correlation ρ_{pb} is a commonly used and well understood effect size measure with some known technical advantages [Coh88, GK05].

To establish a relation between estimates of technically more suitable ρ_{pb}, and more intuitive ω and π, we present the following theorem, showing that all three can be used interchangeably (the proof is given in the appendix). For observations X_i of a random variable X, by S_X^2 we denote $\sum(X_i - \overline{X})^2$ where $\overline{X}$ is an average of observations X_i.

Theorem 1. *Let T^1 and T^2 be two samples of two random variables τ^1 and τ^2. Let X_i be the i-th element in the sorted pooled sample, R_i its rank in that sample, Y_i the corresponding indicator variable, and r_{pb} the sample point biserial correlation between R_i and Y_i. Then, if there are no ties in uncensored data and the censoring time is unique, the following relation holds*

$$W_G = r_{pb} S_R S_Y / |T^1||T^2| \tag{1}$$

Additionally, if $|T^1|/|T^2|$ approaches finite positive constant when $|T^1| \to \infty$,

$$var(W_G) \to var(r_{pb}) S_R^2 S_Y^2 / |T^1|^2 |T^2|^2 \tag{2}$$

also holds when $|T^1| \to \infty$.

Note that the assumptions of the theorem are fulfilled in the context of SAT solving.

We say that two solvers perform the same on one instance if $\rho_{pb} = 0$, or if r_{pb} is not significantly different from 0 in sense of statistical testing. Also, for the measure of difference

	Industrial				Graph coloring			
	A	B	C	D	A	B	C	D
A	-	-0.097	-0.249	-0.229	-	0.206	0.453	0.461
B	0.097	-	-0.241	-0.208	-0.206	-	0.327	0.333
C	0.249	0.241	-	0.072	-0.453	-0.327	-	-0.001
D	0.229	0.208	-0.072	-	-0.461	-0.333	0.001	-

Table 1: Estimates of ρ_{pb} when comparing various solvers. Following labels are used A = MiniSAT 09z, B = minisat cumr r, C = minisat2, D = MiniSat2hack.

d_i between samples of random variables τ_i^1 and τ_i^2 we can take r_i — the estimate of ρ_{pb} for F_i. Statistical significance testing based on r_{pb} values is usually done after the Fisher transformation $z = \frac{1}{2}\log\frac{1+r}{1-r}$. To check the statistical significance of the overall test, for each r_i, value $z(r_i)$ is computed, and those values are averaged. Since all the $z(r_i)$ are asymptotically normally distributed, it is easy to see (using the properties of the normal distribution and asymptotics) that the average $\overline{z}$ is also asymptotically normally distributed:

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

where ρ_i is the population parameter estimated by r_i. To see if the null hypothesis $\overline{\delta} = 0$ holds, one should check if the difference of obtained average $\overline{z}$ from $z(\overline{\delta}) = 0$ is statistically significant with respect to distribution of $\overline{z}$. The p value (two tailed) is $2(1 - \Phi(\overline{z}/\sqrt{var(\overline{z})}))$, where Φ is the distribution function of standard normal distribution. Note that we don't directly use the distribution Θ of $\overline{d}$ because the use of transformed values is more reliable.

The estimate of the effect size $\overline{d}$ is the average of values r_i or values $\frac{1-W_G}{2}$ which estimate the probabilites.

3 Experimental Results

The experiments on industrial and graph coloring instances indicate that the values of r_{pb} and W_G sabilize when the number of shuffled variants is around 10 to 15. More about these experiments can be found in [Nik10]. In Table 1 we present estimates of ρ_{pb} for comparisons of each pair of solvers using 15 shuffled variants. The obtained results are not surprising with respect to the results of MiniSat hack track. Only the ordering of minisat2 and MiniSat2hack is different. In all the comparisons the p values (two tailed) are less than 0.001 except when comparing original MiniSAT version and MiniSat2hack on graph coloring instances when it is 0.945. Nevertheless, note that some statistically significant differences can be considered negligible since r_{pb} values are too small (namely, less than 0.1). The ranking is easy to establish. It is ABDC on industrial and CDBA on graph coloring instances, where the same labels are used as in Table 1.

4 Related Work

There are already several papers concerning the comparison of SAT solvers. Le Berre and Simon recognize the importance of this question [LS04]. Also, the possibility that shuffling can change the order of solvers was noticed. It is suggested that the corpora could include shuffled

variants of formulae. On the other hand, this paper is concerned with the usual way of solver comparison. Audemard and Simon further analyze the impact of the shuffling on the number of solved formulae, and conclude that it can be large [AS08].

Etzoni and Etzoni propose the use of statistical tests for censored data for evaluation of speedup learning systems, but the comparison of runtime distributions of instances is not discussed in their context [EE94]. Brglez et al. stress the importance of statistical approach for SAT solver comparison [BLS05, BO07]. Also the importance of runtime distributions for SAT solver comparison is recognized. Statistical tests are used to compare performances of two solvers, but only on one instance. Full methodology that could use a corpus of instances and combine results of testing on individual instances is not devised. Moreover, we exploit the notion of the effect size which is important for such methodology and propose the extension to ranking several solvers using method which takes the nontransitivity issue into account.

Pulina gives an excellent empirical analysis of ranking methods for systems used in automated reasoning and more importantly establishes reasonable properties that those ranking methods should possess [Pul06].

5 Conclusions

We demonstrated that comparison methods that are widely used can be unreliable, and depend on variable naming, ordering of clauses and literals, and random seeds used (see Sect. 1). A new, statistically founded, methodology is proposed for comparison of SAT solvers. It is based on the comparison of runtime distributions instead of single solving times and uses standard effect size measures to quantify the difference between those distributions.

It is found that the needed number of shuffled variants to estimate the effect size between solvers is around 10 to 15. The testing corpora could be somewhat reduced to compensate for this increase of solving time, thus trading some benchmarks for thorough analysis. We regard this approach better, since the results presented by Nikolić do not suggest that the use of large corpora eliminates the significant chance effects on number of solved formulae [Nik10]. The new methodology is able to practically eliminate the chance effects from the comparison (up to p value) and provide information on statistical significance and effect size in the way usual for statistical testing which standard approach does not.

References

[AS08] G. Audemard and L. Simon. Experiments with Small Changes in Conflict-Driven Clause Learning Arghorithms. In *Proc. of the 14th International Conf. on Principles and Practice of Constraint Programming*, 2008.

[BLS05] F. Brglez, X. Y. Li, M. Stallmann. On SAT Instance Classes and a Method for Reliable Performance Experiments with SAT Solvers. *Annals of Mathematics and Artificial Intelligence*, 2005.

[BO07] F. Brglez and J. Osborne. Performance Testing of Combinatorial Solvers With Isomorph Class Instances. In *ECS'07: Experimental Computer Science on Experimental Computer Science*, 2007.

[Coh88] J. Cohen. Statistical Power Analysis for the Behavioral Sciences. *Lawrence Erlbaum Associates*, 1988.

[EE94] O. Etzoni and R. Etzoni. Statistical Methods for Analyzing Speedup Learning Experiments. *Machine Learning*, 1994.

[Geh65] E. Gehan. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika*, 1965.

[GK05] R. Grissom, J. Kimm. Effect Sizes for Research: A Broad Practical Approach. *Lawrence Erlbaum Associates*, 2005.

[LS04] D. Le Berre and L. Simon. The Essentials of the SAT 2003 Competition. In *Theory and Applications of Satisfiability Testing*, 2004.

[Man67] N. Mantel. Ranking Procedures for Arbitrarily Restricted Observations. *Biometrics*, 1967.

[Nik10] M. Nikolić Statistical Methodology fro Comparison of SAT Solvers *SAT 2010*, to appear.

[Pul06] L. Pulina. Empirical evaluation of Scoring Methods In *Proc. of the 3rd European Starting AI Researcher Symposium*, 2006.