Statistical Methodology for Comparison of SAT Solvers*

Mladen Nikolić

Faculty of Mathematics, University of Belgrade,
Belgrade, Studentski Trg 16, Serbia
nikolic@matf.bg.ac.rs

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. Solver comparisons as a method for detecting good ideas are widely recognized in the SAT community. This is the main purpose of competitions of SAT solvers.[1] Their importance is growing, especially because of the significant number of new ideas and solvers that appear each year. Nevertheless, main responsibility for evaluation of new ideas is on the researchers themselves.

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. Four solvers were chosen from the MiniSAT hack track of the SAT 2009 competition — the first, the last, the baseline and one of the medium performance according to the results of the track.[2] We used two benchmark sets. The first consisted of 292 industrial instances used at the MiniSAT hack track and the second of 300 graph coloring instances from the SAT 2002 competition. Each solver was run on 50 shuffled

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

[1] http://www.satcompetition.org/, http://baldur.iti.uka.de/sat-race-2008/

[2] http://www.cril.univ-artois.fr/SAT09/results/ranking.php?idev=25

variants of each benchmark (obtained by reordering the clauses, literals in each clause, and renaming the variables) with cutoff time of 1200 seconds.

First we checked how much the number of solved formulae can vary. A solver was "lucky" if for each formula it was given the shuffled variant that it solves in the shortest time. The solver was "unlucky" if for each formula it was given the shuffled variant that it solves in greatest time (unsolved if such variant exists). For each benchmark set and each solver, results for both the "lucky" and the "unlucky" case are presented in Table 1. For industrial formulae, the number of formulae solved in their original form is also given. The graph coloring instances were already shuffled, so we don't give such information for them. One can see from the table that the variation of the number of solved formulae can be large.

	Industrial			Graph coloring	
Solver	Lucky	Original	Unlucky	Lucky	Unlucky
MiniSAT 09z	161	142	111	180	157
minisat_cumr r	156	139	107	190	180
minisat2	141	121	93	200	183
MiniSat2hack	144	121	93	200	183

Table 1. Number of solved instances for "lucky" and "unlucky" case of each solver.

Second, we investigated the effect of this variation on solver comparison. We checked that for each two solvers, on the industrial instances it is possible to suitably select shuffled variants of each instance to make one benchmark set on which the first solver is better than the second, and another on which the second is better than the first (in this case, both solvers are run on the same shuffled variant of each formula). However, the probability of such event should be also taken into the consideration. For each pair of solvers we performed 10000 simulated pairwise comparisons with shuffled variants chosen on random for each formula in order to estimate the probabilities of each solver in the comparison being the winner. For most of the pairs, changing the outcome of the comparison turned out to be very unlikely. However, when comparing MiniSAT 09z and minisat cumr r on industrial instances the odds of winning are 92% to 8%, when comparing minisat2 and minisat2hack on industrial instances the odds are 6% to 94%, and when comparing minisat2 and minisat2hack on graph coloring instances the odds are 74% to 26%. It is interesting to notice that on industrial instances, the solver that appears to be the best, can be beaten in practice as a result of chance. Also, ordering of minisat2 and minisat2hack would be different from the one obtained at the competition in most of the cases.

Sometimes the use of shuffling is disputed. Its use is not essential for the methodology that will be proposed. The purpose of shuffling is to make a solver choose different paths of the solving process on different runs, and thus obtain information about its runtime distribution. Such an effect could also be achieved without shuffling by changing the random seed the solver uses, and we certainly don't prefer some random seeds to the others. We also performed the similar experiment with random seeds instead of shuffling. The "lucky version" of MiniSAT solved 144 instances, and the "unlucky" one solved 96, which is close to the results obtained by shuffling. Note that the use of randomization is a common practice in modern SAT solvers.

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 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 methodology we propose was conceived for detection of improvements over some base solver, but it can be used without limitation to comparison of two arbitrary solvers. Also, it will be shown how it can be extended for ranking of several solvers. This methodology *is not* concerned with selection of benchmarks. One should choose the benchmarks representative for the problems of interest.

The rest of the paper is organized as follows. In Sect. 2, a brief information on relevant concepts and techniques is given. The proposed methodology is described in Sect. 3 and the experimental results are given in Sect. 4. In Sect. 5, related work is discussed. In Sect. 6 final conclusions are drawn and some directions of possible further work are pointed to. In the appendix, a proof of the theorem from Sect. 3 is given.

2 Preliminaries

In this section we describe concepts and techniques important for understanding the proposed methodology and introduce needed notation.

2.1 Distributions of Solver Running Times

It is well known that solving times for a propositional formula can vary substantially from one solver run to another if the solver uses some random decisions during its work. Also, solving times can change substantially if a syntactical representation of the formula is changed. Distributions of these solving times have been a subject of intensive study [GSCK00,FRV97], resulting in important theoretical insights and understanding of randomized restarts. A runtime distribution of a solver on some instance bears much more information about solver performance than a single run, but it is considerably more expensive to obtain.

2.2 Statistical Hypothesis Testing and the Notion of the Effect Size

Statistical hypothesis testing is concerned with determining if a proposed hypothesis about some populations hold, based on sample data from those populations. The test is performed by trying to reject the *null hypothesis* H_0. H_0 is usually a statement of "no effect" claiming that the effect considered is not present in the populations.

In order to test if H_0 holds, one computes a value t of some test statistic T (depending on the purpose and formulation of the test) with a known probability distribution. The probability of obtaining the computed or more extreme value of the statistic, assuming that H_0 is true, is called a p value. If the p value is less than some predetermined threshold α (usually 0.05), the observed event is considered to be too improbable to be observed if H_0 holds, and the hypothesis H_0 is rejected. Such a result is said to be *statistically significant at the level* α. Otherwise ($p > \alpha$), one cannot reject the hypothesis H_0.

The smaller the p value, the greater the confidence that the observed effect is not obtained by chance. Nevertheless, a small p value is not enough to conclude that the effect is large, because it depends both on the size of the effect and the sample size. Even if the effect is statistically highly significant, it can still be too small to be of any practical importance. In order to measure the magnitude of the underlying effect, an *effect size* has to be computed. There are several standard effect size statistics [Ros91,GK05]. One, commonly used when comparing two samples, is a point biserial correlation (often referred to as Pearson's r) [Ros91].

2.3 Point Biserial Correlation

Point biserial correlation ρ_{pb} between two random variables is the correlation between their outcomes and an *indicator variable* with value 1 for outcomes of the first random variable, and value -1 for the outcomes of the other. Its sample estimate r_{pb} is calculated by the formula:

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

where X_i denote observations from both samples, and Y_i are indicator variables. $\overline{X}$ and $\overline{Y}$ are the means of X_i and Y_i. N is the total number of observations. Quantities ρ_{pb} and r_{pb} have values ranging from -1 to $+1$. Absolute values closer to 1 mean that the distributions of random variables exhibit better separation. Values near 0 indicate great overlapping between distributions.

If there is no information about the distribution of the data, the data are often transformed by ranking — each observation in either sample is replaced by its rank in the sorted sample. If there are tied (equal) observations, each of them is assigned the average rank of the ranks that would be attributed to them. The point biserial correlation calculated on ranked data has different properties to the original statistic and is an instance of the Spearman correlation coefficient [DKS51,DM61].

The estimate r_{pb} is asymptotically normally distributed with the mean ρ_{pb}. The variance of r_{pb} is not easy to determine if the ranking is used and if the distribution of the data is not normal except for the case $\rho_{pb} = 0$ [DKS51,DM61]. Nevertheless, it can be estimated by methods like bootstrapping or jackknife [Efr79,ES81]. The variance of r_{pb} is strongly dependent on value of ρ_{pb}, and r_{pb} is usually used in statistical tests only after the Fisher's variance stabilizing transformation $z(x) = arctanh(x)$ is applied [Hot53]. Also, the transformed variable is much closer to normal distribution than the original one. It has the mean $z(\rho_{pb})$ and its variance can be estimated by $var(r_{pb})(1 - r_{pb}^2)^{-2}$.

In order to interpret the magnitude of r_{pb}, one can follow commonly accepted recommendations by Cohen [Coh88] — effects with $|r_{pb}|$ in the intervals [0,0.1), [0.1,0.3), [0.3,0.5), and [0.5,1], are considered respectively, negligible, small, medium, and large. However, note that these are not strict rules, but rather, reasonable guidelines.

2.4 Accounting for Censored Data

By *censored data* we mean data known to be greater than some threshold value, but of unknown exact value. One well-known test for comparison of two samples which include censored data is the Gehan test [Geh65]. The statistic used in this test can be formulated as follows [Man67]. The *pooled sample* is the sample that includes elements of both samples that are compared. Note that the repetitions of elements are possible. Let U_i be the number of observations in the pooled sample than which the i-th observation in the pooled sample is strictly greater minus the number than which it is strictly less. In the case of unique censoring time, censored observations are treated as equal and greater than all the uncensored observations.[3] Then Gehan statistic is defined by

$$W_G = \frac{1}{|A_1||A_2|} \sum_{i \in A_1} U_i$$

where A_j is a set of indices in the pooled sample of the observations from the j-th sample ($j = 1, 2$). As shown by Gehan [Geh65], using the theory of U statistics [Hoe48,Leh51], Gehan statistic is a consistent estimate of $\omega = P(X > Y) - P(X < Y)$. It is asymptotically normally distributed with the mean ω. The variance of W_G is easy to calculate if $\omega = 0$. In other cases bootstrapping or jackknife estimates can be used [Efr79,ES81]. As in case of r_{pb}, the variance depends on ω, diminishing as ω approaches extreme values -1 or 1.

[3] In the case of varying censoring times, more sophisticated statistics might be used.

3 The Methodology

An overall idea of the proposed methodology for comparing two solvers is simple. For each SAT instance from some benchmark set one should calculate suitably defined difference of performance of two solvers on that instance. If the performances of two solvers are approximately the same for the benchmark set, then the differences on considered instances should mainly cancel out, and the average of the differences couldn't be too large. Note that the concept of runtime distribution is important for our methodology, but in formulation of the methodology we leave the sampling mechanism unspecified. The methodology will be applicable regardless of that choice. First, we outline the methodology, and then, discuss its various aspects.

3.1 The Outline of the Proposed 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.

3.2 Choosing function d

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. Three possible effect size measures will be introduced, and their relations will be analyzed.

Probably the most intuitive indicator of two solvers performing equally on some instance F would be that the probability that the first solver solves the instance in more time than the second solver is equal to the probability that

```
– 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 $\alpha$ ($\alpha < 1$)
– 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 $\tau_i^j$ ($j = 1, 2$)
    • 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 $\Theta$ — the distribution of $\overline{d}$ under the null hypothesis
– Calculate the $p$ value for $\overline{d}$ according to the distribution $\Theta$
– 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
```

Fig. 1. Outline of the proposed methodology.

the second solver solves the instance in more time than the first solver. More formally

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

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. Namely, the case $\tau^1 = \tau^2$ is possible only for censored observations since, practically, all uncensored solving times differ even slightly if measured with enough precision. The value π is a known effect size measure [GK05]. Recall that ω is estimated by W_G and π is estimated by $(1 - W_G)/2$. Drawback of using ω or π is a lack of variance stabilizing transformation like the one available for the point biserial correlation (see Sect. 2).

Point biserial correlation ρ_{pb} is a commonly used and well understood effect size measure (as described in Sect. 2). It is estimated by r_{pb}. Since there is no information about distribution of the data, estimate should be calculated on ranked data (see Sect. 2). Technical advantage of using this measure is availability of Fisher's transformation which stabilizes the variance and makes the distribution closer to normal. This makes determining statistical significance much more reliable. On the other hand, it is not obvious if this measure makes sense with censored data. Also, without prior experience with this measure, one might feel uncomfortable interpreting its magnitude.

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. As already noticed, the assumption of no ties is quite realistic for uncensored data. The assumption of unique censoring time is standard in SAT solving. The last assumption is trivially satisfied as one can always use samples of equal size. This theorem allows us to use either of the proposed effect size measures for function d since one can be easily calculated from the other. Since p value depends on the value of the test statistic and its variance, the second relation ensures that p value estimates are practically the same for large samples regardless which of the proposed measures is used.

For our primary effect size measure, we take point biserial correlation due to its technical advantages concerning the computation of statistical significance, but ω and π can also be reported for the effect size.

3.3 Determining Statistical Significance and the Effect Size

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 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 (see Sect. 2). 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, and its magnitude is interpreted in the way described in Sect. 2.

3.4 Ranking Several Solvers

If one is comparing several solvers, even if all pairwise comparison results are known one still needs a ranking method.

Important issue with application of statistical tests in general is their potential nontransitivity. Namely, there are examples of random variables A, B, and C such that $P(A < B) > \frac{1}{2}$ and $P(B < C) > \frac{1}{2}$ hold, but $P(A < C) > \frac{1}{2}$ does not. Note that this counterintuitive behavior is not a flaw of any test, but rather a natural probabilistic phenomenon. A popular example are Efron's dice [BH02].

There is still no proof that the proposed comparison procedure is transitive. As with Efron's dice it might be even meaningless to demand transitivity, but this should be a subject of a further study. To overcome this difficulty, one can use Kendal-Wei method for ranking based on pairwise comparisons [Ken55]. This method is designed for situations characterized by nontransitivity property.

4 Experimental Results

In this section we present two experiments. The first one is concerned with the number of shuffled variants appropriate for the application of the methodology, and the second one shows results of the application of the methodology. In both experiments we use the same 4 solvers and 2 benchmark sets as in Sect. 1. For the level of statistical significance α we take the usual value of 0.05. We sample from the runtime distributions by solving 50 shuffled variants of each formula with cutoff time of 1200 seconds. Though the shuffling is quite acceptable for the solvers used, one could also change the random seed. If all the shuffled variants of the benchmark were solved in less than 0.1 seconds[4] by both solvers or no shuffled variant was solved by any solver, the benchmark was discarded as uninformative. For function d we choose r_{pb}. The variance of r_{pb} is estimated by bootstrapping [Efr79] with 100000 bootstraps.[5]

First important question concerning the application of the proposed methodology is its computational cost reflected by the number of shuffled variants one has to use in order to obtain reliable estimates of the effect size and statistical

[4] At most 1 industrial and 30 graph coloring instances were discarded in any comparison on the basis of this criterion.

[5] Source code of software used for all the statistical calculations is available from `http://www.matf.bg.ac.rs/~nikolic/solvercomparison/sc.zip`

significance. Also, increasing the number of shuffled variants leads to smaller p values due to larger sample size without the increase of the effect size. It is advised that the sample size is not increased beyond the point at which the effect size estimate stabilizes [Coh95]. To check the number of needed shuffled variants, for each two solvers, we plotted the value of r_{pb} as the number of used shuffled variants ranges from 1 to 50. The plot for each benchmark set is given in Fig. 2. The plots indicate that the number of shuffled variants that should be used is around 10 to 15. As expected, the results of the experiments based on the estimates of ω and π instead of ρ_{pb} are the same.

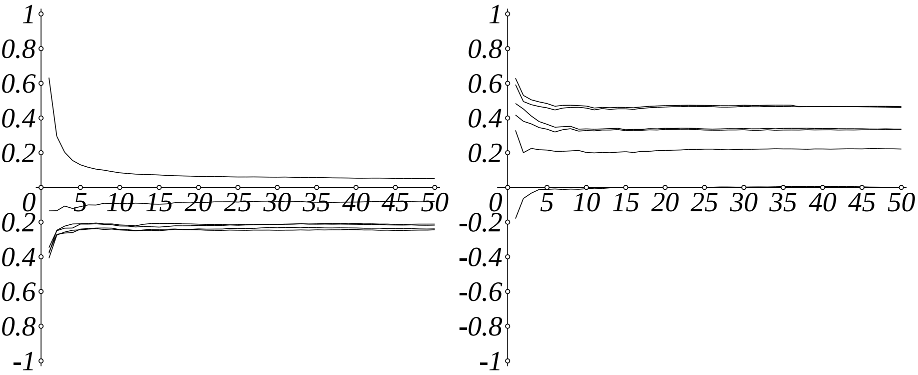

Fig. 2. Plots of r_{pb} for industrial (left) and graph coloring (right) benchmark sets as a function of the number of shuffled variants used.

In Table 2 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 those shown in Table 1. 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 with respect to guidelines provided in Sect. 2. Note that no problems with transitivity appeared. 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 2.

5 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].

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

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].

6 Conclusions and Future Work

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.

We showed 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 in Sect. 1 do not suggest that the use of large corpora eliminates the significant chance effects on number of solved formulae. 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.

As for the future work, important issue is finding the assumptions that guarantee the transitivity of proposed comparison procedure, and checking if nontransitive effects can appear in SAT solving. Also, proposed ranking method is yet to be analyzed in the light of the criteria established by Pulina [Pul06].

Acknowledgements

The author is grateful to Predrag Janičić, Filip Marić, and Oliver Kullmann for their valuable suggestions and discussion about this work.

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.

[BH02] B. Brown and T. Hettmansperger. Kruskal-Wallis, Multiple Comparisons and Efron Dice. *Australian & New Zealand Journal of Statistics*, 2002.

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

[Coh95] P. Cohen. Empirical Methods for Artificial Intelligence. *The MIT Press*, 1995.

[Cra46] H. Cramér. Mathematical Methods of Statistics. *Princeton Univeristy Press*, 1946.

[DM61] F. David and C. Mallows. The Variance of Spearman's rho in normal samples. *Biometrika*, 1961.

[DKS51] S. David, M. Kendall, and A. Stuart. Some Questions of Distribution in the Theory of Rank Correlation. *Biometrika*, 1951.

[Efr79] B. Efron. Bootstrap Methods: Another Look at Jackknife. *The Annals of Statistics*, 1979.

[ES81] B. Efron and C. Stein. The Jackknife Estimate of Variance. *The Annals of Statistics*, 1981.

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

[FRV97] D. Frost, I. Rish, and L. Vila. Summarizing CSP hardness with continuous probability distributions. In *Proc. of the 14th National Conf. on Artificial Intelligence*, 1997.

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

[GSCK00] C. Gomes, B. Selman, N. Crato, H. Kautz. Heavy-Tailed Phenomena in Satisfiability and Constraint Satisfaction Problems. *Journal of Automated Reasoning*, 2000.

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

[Hoe48] W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 1948.

[Hot53] H. Hotelling. New Light on the Correlation Coefficient and its Transforms *Journal of the Royal Statistical Society*, 1953.

[Ken55] M. Kendall. Further Contributions to the Theory of Paired Comparisons. *Biometrics*, 1955.

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

[Leh51] E. Lehmann. Consistency and Unbiasedness of Certain Nonparametric Tests. *The Annals of Mathematical Statistics*, 1951.

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

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

[Ros91] R. Rosenthal. Meta-Analytic Procedures for Social Research. *Sage*, 1991.

[Zar05] E. Zarpas. Benchmarking SAT Solvers for Bounded Model Checking. In *Theory and Applications of Satisfiability Testing*, 2005.

Appendix

Proof of Theorem 1.
Let $n_1 = |T^1|$, $n_2 = |T^2|$, and $N = n_1 + n_2$. The numbers of censored observations in each sample are denoted by c_1 and c_2, and $C = c_1 + c_2$. Let I^1 and I^2 be the sets of indices in the pooled sample of uncensored observations from samples T^1 and T^2 respectively. Let $I = I^1 \cup I^2$. The set of indices in the pooled sample of all the observations of the first sample is denoted by A_1.

First we show that the relation (1) holds. We will consider expressions $n_1 n_2 W_G$ and $S_R S_Y r_{pb}$ and will conclude that they are equal. We use Mantel's version of W_G [Man67] noting that it can be decomposed in terms of ranks of uncensored observations plus the term for censored observations.

$$\begin{aligned} n_1 n_2 W_G = \sum_{i \in A_1} U_i &= \sum_{i \in I^1} [(R_i - 1) - (N - R_i)] + c_1(N - C) \\ &= 2 \sum_{\in I^1} R_i - (n_1 - c_1)(N + 1) + c_1(N - C) \\ &= 2 \sum_{\in I^1} R_i - (n_1 - 2c_1)(N + 1) - c_1(C + 1) \end{aligned}$$

Let us consider $S_R S_Y r_{pb}$:

$$S_R S_Y r_{pb} = \sum_{i=1}^{N} (R_i - \overline{R})(Y_i - \overline{Y}) = \sum_{i=1}^{N} R_i Y_i - \sum_{i=1}^{N} R_i \overline{Y} - \sum_{i=1}^{N} \overline{R} Y_i + \sum_{i=1}^{N} \overline{RY}$$

where $\overline{R}$ and $\overline{Y}$ are the means of R_i and Y_i. Note that the last three sums are equal, and hence

$$S_R S_Y r_{pb} = \sum_{i=1}^{N} R_i Y_i - \sum_{i=1}^{N} R_i \overline{Y} = \sum_{i=1}^{N} R_i Y_i - E_1$$

where $E_1 = (N+1)(n_1 - n_2)/2$ and is obtained using the fact that the sum of ranks is constant and equals $N(N+1)/2$ and that $\overline{Y} = (n_1 - n_2)/N$. Separating censored and uncensored observations yields

$$S_R S_Y r_{pb} = \sum_{i \in I} R_i Y_i + E_2 - E_1 = \sum_{i \in I^1} R_i - \sum_{i \in I^2} R_i + E_2 - E_1$$

where $E_2 = (2N - C + 1)(c_1 - c_2)/2$ since $(2N - C + 1)/2$ is the average rank of the censored observations. Since all the uncensored observations are less than censored ones, and since the sum of their ranks is constant, the second sum can be expressed in terms of the first sum:

$$S_R S_Y r_{pb} = 2 \sum_{i \in I^1} R_i - (N - C)(N - C + 1)/2 + E_2 - E_1$$

After elementary calculations we obtain:

$$S_R S_Y r_{pb} = 2 \sum_{\in I^1} R_i - (n_1 - 2c_1)(N+1) - c_1(C+1)$$

thus proving the relation (1).

To prove the relation (2), we note that S_Y is constant, and that S_R is constant for fixed c_1 and c_2. For convenience, we will talk in terms of ratios $a_1 = c_1/n_1$ and $a_2 = c_1/n_2$. Using (1), the conditional variance of W_G is $var(W_G|a_1, a_2) = \frac{S_R^2 S_Y^2}{n_1^2 n_2^2} var(r_{pb})$. We need to prove $var(W_G)/var(W_G|a_1, a_2) \to 1$ when $n_1 \to \infty$. We will follow the reasoning of Gehan [Geh65]. By the law of total variance we have

$$var(W_G) = E_{l_1,l_2} var(W_G|l_1, l_2) + var_{l_1,l_2} E(W_G|l_1, l_2)$$

By the law of large numbers, a_1 and a_2 converge in probability to their expectations α_1 and α_2 when $n_1 \to \infty$. Since the probabilities of l_i such that $|l_i - \alpha_i| \geq \varepsilon$ vanish for all $\varepsilon > 0$ when $n_1 \to \infty$, it holds

$$\frac{n_1^{-3} E_{l_1,l_2} var(W_G|l_1, l_2)}{n_1^{-3} var(W_G|a_1, a_2)} \to 1$$

when $n_1 \to 1$. The last relation is obtained using the convergence theorems by Cramér and Slutsky [Cra46] which can be used since it is known that $n_1^{-3} var(W_G|a_1, a_2) = O(1)$ when $n_1 \to \infty$ [Geh65].

Regarding the second term in the expansion of unconditional variance, by definition

$$var_{l_1,l_2} E(W_G|l_1, l_2) = E_{l_1,l_2} E^2(W_G|l_1, l_2) - (E_{l_1,l_2} E(W_G|l_1, l_2))^2$$

which converges to 0 by similar reasoning as for the first term. This proves the convergence (2). □