Competitive evaluation of automated reasoning tools: Empirical scoring and statistical testing

Massimo Narizzano, Luca Pulina, and Armando Tacchella*

DIST, Università di Genova, Viale Causa, 13 – 16145 Genova, Italy
{mox,pulina,tac}@.dist.unige.it

Abstract. Empirical scoring is the most common ranking method in automated reasoning systems competitions. Statistical testing can be used to validate the results of scoring, since the null hypothesis of equal performances is tested against the alternative hypothesis of significant difference in performances using a precise mathematical formulation. This paper evaluates the merits of statistical testing as a complement to empirical scoring using the 2005 comparative evaluation of solvers for quantified Boolean formulas as a case study.

1 Introduction

The automated reasoning research community has grown accustomed to competitive events where a pool of systems is run on a pool of problem instances with the purpose of ranking the systems according to their performances. A non-exhaustive list of such events includes the CADE ATP System Competition (CASC) [1], the SAT Competition [2], the International Planning Competition (see, e.g., [3]), and the CP Competition (see, e.g., [4]). The main purpose of such events is to designate a winner. Even if the results of competitions may provide less insight than controlled experiments in the spirit of [5], there is a general agreement that they raise interest in the community and play a fundamental role in the advancement of the state of the art.

In this paper, we consider two different aspects that can play a crucial role in a competition: empirical scoring of the systems, and statistical testing of the alleged results. Empirical scoring is the most common ranking method, whereby a tournament-like procedure is used to assign bonuses and penalties to each system according to various performance indicators. The main advantages of such procedures are their simplicity and their wide applicability, but they offer no direct way of assessing the quality of the rankings provided. Statistical testing, on the other hand, is less commonly used (see, e.g., [3]), usually more complicated, and less widely applicable than empirical scoring. However, statistical testing provides direct means of assessing the result quality, since the null hypothesis of equal performances is tested against the alternative hypothesis of significant difference in performances using a precise mathematical formulation that allows an estimate of the results within some stated confidence level.

Our analysis considers seven scoring methods (references can be found in [6]): three previously used in systems competitions, i.e., CASC, the SAT competition, and the evaluation of solvers for quantified Boolean formulas (QBFs); three adapted from the eponymous voting systems, i.e., Borda count, range voting and Schulze's method; and, finally, YASM ("Yet Another Scoring Method") described, e.g., in [6]. Using the data of the 2005 comparative evaluation of QBF solvers (QBFEVAL'05 [7]) as a case study, we summarize the results presented in [6] to show how YASM provides the best compromise among the scoring methods considered over a set of measures that quantify desirable properties of the methods. Then, we assess the significance of the ranking obatined by YASM using two statistical tests inspired by the approach of [3], namely the Wilcoxon signed rank test (see, e.g., Ch. 12a of [8]) and the Wilcoxon rank sum test (also known as Mann-Whitney test; see, e.g., Ch.11a of [8]). Our goal is to assess whether the ranking obtained with an empirical scoring method like YASM is compatible with the significance results inferred by statistical testing, and whether statistical testing can help us to improve the scientific value of competitions.

*The authors wish to thank the Italian Ministry of University and Research (MIUR) for its financial support, and the anonymous reviewers who helped to improve the original manuscript.

	CASC/QBF	SAT	YASM	Borda	range voting	Schulze
OPENQBF	201	62621.96	482.92	436	1682	2
QBFBDD	106	39250.40	363.74	338	3236	1
QMRES	227	173068.18	1505.03	1085	12050	4
QUANTOR	318	228854.86	2701.35	2019	32393	8
SEMPROP	289	148690.91	1787.92	1569	18317	7
SSOLVE	243	110121.36	1415.36	1286	16038	6
WALKQSAT	189	87535.25	1090.98	962	11010	3
YQUAFFLE	250	110257.07	1351.92	1200	11864	5

Table 1. Scores of the solvers (listed in alphabetical order) according to the methods considered. Schulze's scores are the straightworward translation of the ordinal ranking derived by applying the method which is not based on cardinal scoring.

2 Preliminaries

The results of QBFEVAL'05 can be listed in a table RUNS comprised of four attributes: SOLVER, INSTANCE, RESULT, and CPUTIME. The attributes SOLVER and INSTANCE report which solver is run on which instance. RESULT is a four-valued attribute: SAT (resp. UNSAT), i.e., the instance was found satisfiable (resp. unsatisfiable) by the solver, TIME, i.e., the solver exceeded the time limit (900 seconds), and FAIL, i.e., the solver aborted for some reason beyond our control. Finally, CPUTIME reports the CPU time spent by the solver on the given instance (see [7] for more details).

The analysis herewith presented rests on the assumption that a table identical to RUNS is the only input required by a scoring method. As a consequence, we do not take into account (*i*) memory consumption, (*ii*) correctness of the solution, and (*iii*) "quality" of the solution. The only measures of merit at our disposal are the number of problems solved and the CPU time of the solvers. Notice that the number of problems solved is correct as long as the CPU time measure used to enforce the time limit is so. Therefore, to ensure accuracy of the empirical scoring methods it is very important to tame potential sources of errors in the CPU time measures. A detailed analysis of such errors, and a possible solution to deal with them, have been presented in [6]. Here we observe that statistical testing naturally obviates to this issue, since the samples are already assumed to be noisy observations of the true (unknown) values.

We conclude our preliminaries by briefly describing the empirical scoring methods used in our analysis (references and further details provided in [6]). For the sake of comparison, Table 1 shows the total scores earned by QBFEVAL'05 participants according to the methods considered.

CASC Solvers are ranked according to the number of times that RESULT is either SAT or UNSAT. In case of a tie, solvers faring the lowest CPUTIME averaged over the problems solved are preferred.

QBF evaluation QBFEVAL scoring method is the same as CASC, except that ties are broken using CPUTIME summed over the problems solved.

SAT competition The last SAT competition uses a purse-based method, i.e., the score of a solver is obtained by adding up three purses: the solution purse is divided equally among all solvers that solve a problem; the speed purse is divided unequally among all the competitors that solve a problem according to their relative speed; finally, the series purse is divided equally among all the solvers that solve at least one problem in a family of related instances (series).

Borda count Given n solvers, each voter (instance) ranks the candidates (solvers) in ascending order considering the value of the CPUTIME field. Let $p_{s,i}$ be the position of a solver s in the ranking associated with instance i ($1 \leq p_{s,i} \leq n$). The score of s is $S_{s,i} = n - p_{s,i}$. In case of time limit attainment and failure, $S_{s,i} = 0$. The total score S_s of a solver s is the sum of all the scores $S_{s,i}$.

Range voting Similar to Borda count, whereas an arbitrary scale is used to associate a weight w_p with each of the n positions and the score $S_{s,i}$ is computed as $S_{s,i} = w_p \cdot p_{s,i}$ (default to 0 in case of time limit attainment or failure).

Schulze's method We denote as such an extension of the method described in Appendix 3 of [9]. Since Schulze's method is meant to compute a single overall winner, we extended the method

according to Schulze's suggestions [10] in order to make it capable of generating an overall ranking.

YASM This method (named YASMv2 in [6]) is influenced by three factors, namely (*i*) a Borda-like positional weight (*ii*) the relative hardness of the instances, and (*iii*) the relative speed of the solver with respect to the fastest solver on the instance.

3 Assessment of Empirical Scoring

In this section we summarize the results obtained considering the above mentioned scoring methods and the following effectiveness measures.

Homogeneity The rationale behind this measure is to verify that, on a given test set, the scoring methods considered (*i*) do not produce exactly the same solver rankings, but, at the same time, (*ii*) do not yield antithetic solver rankings. We compute homogeneity with the Kendall rank correlation coefficient τ and, considering QBFEVAL'05 data, only two pairs of methods (QBF-CASC and Schulze-Borda) show perfect agreement ($\tau = 1$), while all the other pairs agree to some extent, but still produce different rankings (see [6] for the homogeneity table).

Fidelity We proposed fidelity in [6] to check whether the scoring methods introduce any distortion with respect to the true merits of the solvers. Of course, we have no way to know such merits. Thus we measure fidelity by feeding each scoring method with "white noise", i.e., a table RUNS filled with random results so we know in advance that the (synthesized) merit of the competitors is approximately the same, and we can measure distortions with respect to this pattern. In [6] we consider the percent ratio F between the lowest score and the highest one: a higher value of F implies higher fidelity. In [6] we show that YASM, yielding the highest value of F among all the scoring methods, is the method with the highest fidelity.

RDT-, DTL-, and SBT-stability Stability on a randomized decreasing test set (RDT-stability), and stability on a decreasing time limit (DTL-stability) are meant to measure how much a scoring method is sensitive to perturbations that diminish the size of the original test set, and how much a scoring method is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers, respectively. However, in [6] we conclude that these two measures do not help to discriminate among the scoring methods. Stability on a solver biased test set (SBT-stability), measures how much a scoring method is sensitive to a test set that is biased in favor of a solver. A high SBT-stability coupled with a high fidelity is a good indicator that the method is able to detect the absolute merit of the solvers. YASM performance in terms of SBT stability lies in between CASC/QBF and SAT, on one side, and voting systems, on the other side. In [6] we report that YASM turns out to be, on average, better than CASC/QBF and SAT, while it is worse, on average, than the methods based on voting systems. However, YASM offers the best compromise between SBT-stability and fidelity.

SOTA-relevance This measure (see, e.g., [6]) is meant to capture the relationship between the ranking obtained with a scoring method and the strength of a solver, as witnessed by its contribution to the SOTA solver, i.e., the ideal solver that always fares the best time among all the participants. The results presented in [6], show that CASC/QBF methods turn out to have the highest SOTA-relevance ($\tau = 1$), while YASM is only third best with $\tau = 0.79$.

4 Empirical Scoring and Statistical Testing

As anticipated in Section 1, we have no direct means of assessing the significance of the results obtained by an empirical scoring method like YASM. Indeed, even if we can do this indirectly using the measures presented in the previous Section, there is no guarantee that the results obtained will apply to a different set of solvers and/or problem instances. On the other hand, if we rephrase the problem in terms of statistical hypothesis testing, then we can check for statistically significant differences in the performances of the solvers and validate our conclusions within some stated

confidence level. Let us start by introducing a null hypothesis and an alternative hypotheses that are appropriate in our context. Given any two solvers A and B we can state the:

null hypothesis (H_0), i.e., there are no significant differences in the performances of A with respect to the performances of B; and the

alternative hypothesis (H_1), i.e., there are significant differences in the performances of A with respect to the performance of B.

In the following, let X_A and X_B be the vectors of run-time values associated to solver A and solver B, respectively. Before applying statistical methods to QBFEVAL'05 data, we must decide (i) how to consider missing values, i.e., TIMEOUT and FAIL values, and (ii) which assumptions, if any, can be made about the run-time distributions. The above issues have an impact over the specific method that we can apply to test H_0, because some methods cannot deal with missing values in X_A and X_B seamlessly, and most methods require binding assumptions about the underlying distribution of X_A and X_B.

Considering QBFEVAL'05 data, after removing the instances where all the solvers either fail or reach the time limit, there are two possible models to deal with the remaining missing values: failure-as-time-limit (FAT) model, and time-limit-as-failure (TAF) model. In the FAT model, each time that a solver fails or exceeds the time limit, we default its run time to the time limit. This model (used, e.g., in [11]) consistently overestimates the performances of the solvers, but allows the paired comparison of the values in X_A and in X_B. In the TAF model, each time a solver fails or exceeds the time limit, we simply disregard the data point. In this way overestimation does not occur, but since the vectors X_A and X_B may not be equal in length, the paired comparison of run-times is not generally possible.

Given FAT and TAF data models, we may ask whether an underlying normal distribution of run-times can be assumed. If so, well-known classical techniques like paired t-tests or Analysis of Variance (ANOVA) [12] could be used to test for H_0. Thus, for each solver A, we check X_A under FAT and TAF models using the Shapiro-Wilk [12] test of the null hypothesis that the X_A's are obtained from a normally distributed population. All such tests yield p-values in the order of 10^{-27} for the FAT model, and of 10^{-24} for the TAF model, indicating that it is highly unlikely that the run-time distribution of some solver is anywhere close to normal.[1] Because of this, we must resort to non-parametric tests. Inspired by [3], we consider the Wilcoxon signed rank (WSR) test, a non-parametric alternative to the classical paired t-test, whereby the null hypothesis states that X_A and X_B do not differ in a significant way (see, e.g., Ch. 12a of [8]). The WSR test is applicable as long as:

1. the paired values of X_A and X_B are randomly and independently drawn;
2. the dependent variable (i.e., the run-time) is intrinsically continuous; and
3. it makes sense to compare the values in X_A and X_B.

Both FAT and TAF models fulfill the above conditions, but considering the mechanics of the WSR test, we see that it requires the vector $X_A - X_B$ to be computed. Therefore, it can be applied only in the context of the FAT model. In order to cope with the TAF model, we consider the Wilcoxon rank sum test, also known as Mann-Whitney test (see, e.g., Ch. 12a of [8]). Such test, that we call hereafter Mann-Whitney-Wilcoxon (MWW) test, is a non-parametric test of difference between X_A and X_B, and it can accommodate for unequal lengths of X_A and X_B. In particular, MWW is applicable as long as conditions 2 and 3 above hold, and it requires that the values of X_A and X_B are independently drawn: since in our case the data are (positively) dependent, MWW gives approximate, although conservative, solutions. In the following, all the data and the results extrapolated from the WSR and MWW tests are implicitly referred to the FAT and TAF models, respectively.

[1] We performed all the statistical computations described in this Section using R [13]. Our data tables and R macros are downloadable from `http://www.star.dist.unige.it/~tac/Publications/2006_EMAA.tar.gz`.

	openQBF		qbfbdd		QMRes		quantor		semprop		ssolve		WalkQSAT	
	WSR	MWW	WSR	MWW	WSR	MWW	WSR	MWW	WSR	MWW	WSR	MWW	WSR	MWW
qbfbdd	<0.001	<0.001	–	–	–	–	–	–	–	–	–	–	–	–
QMRes	0.003	<0.001	**<0.001**	**1.000**	–	–	–	–	–	–	–	–	–	–
quantor	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	–	–	–	–	–	–	–	–
semprop	<0.001	<0.001	**<0.001**	**0.982**	<0.001	0.020	<0.001	0.003	–	–	–	–	–	–
ssolve	<0.001	<0.001	**<0.001**	**0.982**	0.059	0.024	<0.001	0.002	0.031	1.000	–	–	–	–
WalkQSAT	<0.001	<0.001	**<0.001**	**0.161**	**0.521**	**<0.001**	**<0.001**	**0.566**	**<0.001**	**0.982**	<0.001	0.982	–	–
yQuaffle	<0.001	<0.001	**<0.001**	**1.000**	0.018	0.057	<0.001	<0.001	**<0.001**	**1.000**	0.521	1.000	**<0.001**	**0.646**

Table 2. p-values of Wilcoxon signed rank (WSR) and Mann-Whitney-Wilcoxon (MWW) tests.

The results of the pairwise WSR and MWW tests on QBFEVAL'05 data are shown in Table 2. In the Table, for each pair of solvers A (row) and B (column) we report the p-value adjusted for multiple comparisons of the pairwise WSR tests and pairwise MWW tests. Before analyzing the results of Table 2 it is worth mentioning that adjustment of the p-values is necessary, because performing multiple comparisons raises the risk of obtaining positive results just by the effect of chance (a fact that follows from Bonferroni inequalities [14]). We applied Holm's adjustment method that, according to [13], is more effective than the well known Bonferroni's method [14]. Both methods give strong control on the family wise error rate, i.e., the probability that the number of overall false rejections (false positives) is greater or equal to one. In [3] the authors use a different correction, i.e., they compute the overall α_0 using $\alpha_0 = 1 - (1 - \alpha)^{(1/n)}$, where α is the confidence level of each test and n is the number of tests performed. However, in [14], the latter kind of correction is said to be less generally valid than Bonferroni's (and thus also Holm's) method. With this proviso, since we wish to extrapolate a partial order of the solvers in terms of their relative speed performances at a 99% confidence level, we reject H_0 only when the p-value shown in Table 2 is less than 0.01. By looking at Table 2 we can see that the WSR test finds most pairwise comparisons significant at an *overall* 99% confidence level. On the other hand, the MWW test yields more conservative results, i.e., 9 comparisons that are significant for the WSR test fail to be so for the MWW test, and only 1 comparison that is not significant for the WSR test is indeed significant for the MWW test. In Table 2, we highlight in boldface the cases in which the two tests are found disagreeing about the significance of the difference in terms of performances. Notice that WSR and MWW tests are more sensitive to a consistent, albeit small, difference in performances rather than occasional, albeit large, differences.

In Figure 1, we represent two partial orders of the solvers in terms of their performances derived from QBFEVAL'05 data and the results of Table 2. The partial order on the left of Figure 1 is based on the results of the WSR test, while the one on the right is based on the results of the MWW test. Both partial orders are obtained by drawing an edge from A to B whenever there is a significant difference in performances between A and B, while the direction of the edge is obtained considering the pairs (x_A, x_B) such that $x_A \in X_A$ and $x_B \in X_B$ and computing the ratio $R(A, B)$ between (i) the number of pairs such that $x_A < x_B$ and (ii) the number of pairs such that $x_B > x_A$ (ties are thus excluded): if $R(A, B) > 1$, then A is faster than B. In Figure 1 we label each directed edge from A to B with the value of $R(A, B)$ and we omit the links that can be extrapolated by transitive closure. The calculation of R is inspired by the WSR test mechanics and, as such, it is meant to reflect consistent differences in performances rather than occasional large gaps. Looking at Figure 1 we can see that the partial order induced by the results of the WSR test is compatible with the one induced by the results of the MWW test, although the latter is less constrained than the former.

We can now compare the snapshots of QBFEVAL'05 data offered by YASM and the other empirical scoring methods (Table 1) with the ones offered by statistical testing (Figure 1). The first observation is that all the rankings produced by the scoring methods are compatible with the partial orders of Figure 1 with the only exception of SAT, which ranks QMRes above semprop, while there is a consistent and significant difference in performances detected by the WSR test that, together with the value of $R(\text{semprop}, \text{QMRes})$, prompts us otherwise. However, notice

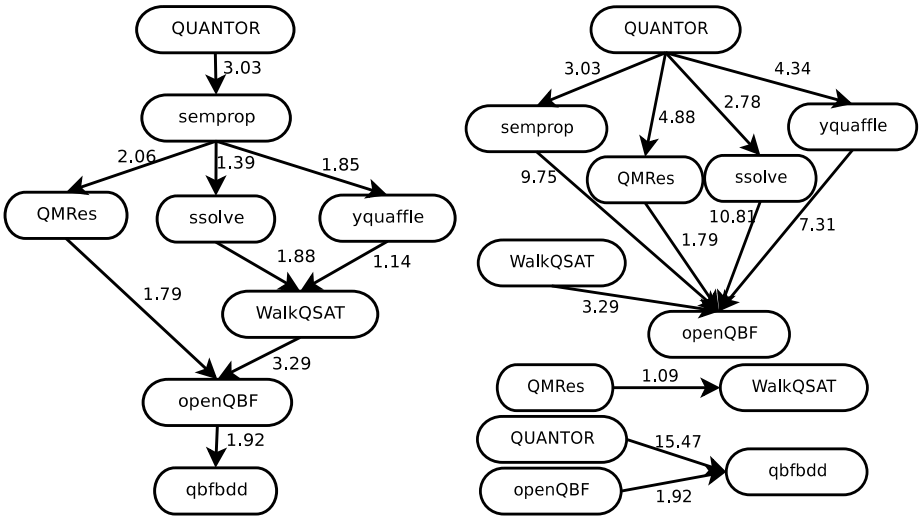

Fig. 1. Partial order of the solvers in terms of their relative speed performances extrapolated from the results of Wilcoxon signed rank (left) and Mann-Whiteny-Wilcoxon (right) tests.

that according to the MWW test, such difference is not significant at the 99% confidence level. The second observation is that, looking at Figure 1 we can see that QMRes, ssolve and yQuaffle are found essentially "incomparable" by the WSR test, and, indeed, the ranking of such solvers is essentially the only part where the scoring methods differ. In particular, if we consider the relative performance index offered by $R(\text{SEMPROP}, B)$, where B is one of QMRes, ssolveand yQuaffle, we can see that only Borda count and Schulze's method rank the three solvers according to the reverse order of the corresponding edge labels. On the other hand, according to the MWW test (Figure 1, right), also semprop is incomparable to the above three solvers, but semprop always ranks second best according to all the scoring methods (with the above mentioned exception of SAT). Finally, let us consider the total orders extracted by the partial ones of Figure 1 by proceeding top-down and breaking the ties considering the edge labels in reverse order. If we compare the total orders thereby obtained with those resulting from Table 1 using the Kendall coefficient, then we can observe the following. Comparing the WSR- and MWW-based total orders yields $\tau = 0.93$, an almost perfect agreement tainted only by the different classification of semprop and ssolve. Considering the empirical scoring methods, it turns out that the WSR-based total order yields $\tau = 1$ in the case of Borda count and Schulze's method, and $\tau < 1$ in all the other cases, with YASM being the closest (together with range voting) at $\tau = 0.86$. In the case of the MWW-based total order, $\tau < 1$ for all the empirical scoring methods considered: YASM, with $\tau = 0.79$ is closer than both SAT ($\tau = 0.64$) and CASC/QBF ($\tau = 0.76$), but range voting ($\tau = 0.79$), Schulze's method and Borda count (both at $\tau = 0.93$) are even closer.

5 Conclusions

Summing up, the analysis presented in this paper confirmed that statistical tests can provide an essential complement to any empirical scoring method in order to check whether the rankings thereby obtained are not due to chanche alone. However, we believe that the widespread adoption of statistical testing in the automated reasoning community has to be supported by further investigation. In particular, it is not clear whether the current statistical tools are adequate for the kind of distributions that can be ascribed to hard combinatorial problems, and whether such tools can scale smoothly to handle the large amount of data involved in the running of competitions or other empirical investigations concerning automated reasoning systems.

References

1. G. Sutcliffe and C. Suttner. The CADE ATP System Competition. `http://www.cs.miami.edu/~tptp/CASC` [2006-6-2].
2. D. Le Berre and L. Simon. The SAT Competition. `http://www.satcompetition.org` [2006-6-2].
3. D. Long and M. Fox. The 3rd International Planning Competition: Results and Analysis. *Artificial Intelligence Research*, 20:1–59, 2003.
4. M.R.C. van Dongen. Introduction to the Solver Competition. In *CPAI 2005 proceedings*, 2005.
5. J. N. Hooker. Testing Heuristics: We Have It All Wrong. *Journal of Heuristics*, 1:33–42, 1996.
6. M. Narizzano, L. Pulina, and A. Tacchella. Competitive Evaluation of QBF Solvers: noisy data and scoring methods. Submitted to LPAR 2006. Draft available on-line at `http://www.star.dist.unige.it/~tac/Publications/2006-LPAR_NarPulTac.pdf` [2006-5-9].
7. M. Narizzano, L. Pulina, and A. Tacchella. The third QBF solvers comparative evaluation. *Journal on Satisfiability, Boolean Modeling and Computation*, 2:145–164, 2006. Available on-line at `http://jsat.ewi.tudelft.nl/` [2006-6-2].
8. R. Lowry. *Concepts and Applications of Inferential Statistics*. Vassar College, 2006. Available on line at `http://faculty.vassar.edu/lowry/webtext.html` [2006-6-2].
9. M. Schulze. A New Monotonic and Clone-Independent Single-Winner Election Method. *Voting Matters*, pages 9–19, 2003.
10. M. Schulze. Extending schulze's method to obtain an overall ranking. Personal communications.
11. J. N. Hooker and V. Vinay. Branching Rules for Satisfiability. *Journal of Automated Reasoning*, 15:359–383, 1995.
12. G.K. Kanji. *100 Statistical Tests - New Edition*. SAGE Publications, 1999.
13. R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2006. Version 2.3.1 (2006-06-01) - Available from `http://cran.r-project.org/doc/manuals/fullrefman.pdf`.
14. E. W. Weisstein. Bonferroni correction. From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/BonferroniCorrection.html` [2005-6-27].