Voting Systems and Automated Reasoning: the QBFEVAL Case Study

Massimo Narizzano, **Luca Pulina** and Armando Tacchella

STAR-Lab
University of Genoa, Italy

COMSOC 2006, Amsterdam, December 6-8

Introduction

- The automated reasoning research community has grown accustomed to competitive events.

Introduction

- The automated reasoning research community has grown accustomed to competitive events.
- An (incomplete) list:
  - CADE ATP System Competition (CASC)
  - SAT Competition
  - QBF Evaluation
  - International Planning Competition
  - ...

Introduction

- The automated reasoning research community has grown accustomed to competitive events.
- An (incomplete) list:
  - CADE ATP System Competition (CASC)
  - SAT Competition
  - QBF Evaluation
  - International Planning Competition
  - ...
- Fundamental role in the advancement of the state of the art:
  - for developers: help to set research challenges
  - for users: assess the current technological frontier

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.
- The ranking should be a representation of the relative strength of the systems.

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.
- The ranking should be a representation of the relative strength of the systems.
- Two sets of aggregation procedures:

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.
- The ranking should be a representation of the relative strength of the systems.
- Two sets of aggregation procedures:
  - methods used in automated reasoning systems contests and a new method called YASM ("Yet Another Scoring Method")

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.
- The ranking should be a representation of the relative strength of the systems.
- Two sets of aggregation procedures:
  - methods used in automated reasoning systems contests and a new method called YASM ("Yet Another Scoring Method")
  - procedures based on voting systems

Introduction

- The competion winner is the system ranking above the others according to some aggregation procedure.
- The ranking should be a representation of the relative strength of the systems.
- Two sets of aggregation procedures:
  - methods used in automated reasoning systems contests and a new method called YASM ("Yet Another Scoring Method")
  - procedures based on voting systems
- We introduce measures to quantify desirable properties of the aggregation procedures.

Contribution

Using and evaluating social choice methods in automated reasoning systems contests

Agenda

- Preliminaries
- Procedures
- YASM
- Comparative measures
- Conclusions

Preliminaries

- Empirical analysis based on QBFEVAL 2005 data:
  - eight solvers of the second stage
  - fixed structure QBF instances

Preliminaries

- Empirical analysis based on QBFEVAL 2005 data:
  - eight solvers of the second stage
  - fixed structure QBF instances
- Table RUNS with four attributes: SOLVER, INSTANCE, RESULT, and CPUTIME.

Preliminaries

- Empirical analysis based on QBFEVAL 2005 data:
  - eight solvers of the second stage
  - fixed structure QBF instances
- Table RUNS with four attributes: SOLVER, INSTANCE, RESULT, and CPUTIME.
- RUNS is the only input required by an aggregation procedure.

Agenda

- Preliminaries
- Methods
- YASM
- Comparative measures
- Conclusions

Procedures used in automated reasoning systems contests

- **CASC:** solvers are ranked according to the number of problems solved and ties are broken using average CPUTIME.

Procedures used in automated reasoning systems contests

- **CASC:** solvers are ranked according to the number of problems solved and ties are broken using average CPUTIME.
- **QBF evaluation:** is the same as CASC but ties are broken using total CPUTIME.

Procedures used in automated reasoning systems contests

- **CASC:** solvers are ranked according to the number of problems solved and ties are broken using average CPUTIME.
- **QBF evaluation:** is the same as CASC but ties are broken using total CPUTIME.
- **SAT competition:** uses a purse-based method where the score is obtained adding up a solution purse, a speed purse and a series purse.

Procedures based on voting systems

Assuming solvers as candidates to an election and instances as voters:

- **Borda count:** solvers are ordered by CPUTIME and to each position is associated a score.

Procedures based on voting systems

Assuming solvers as candidates to an election and instances as voters:

- **Borda count:** solvers are ordered by CPUTIME and to each position is associated a score.
- **Range voting:** similar to Borda count, but using multiplicative positional weights.

Procedures based on voting systems

Assuming solvers as candidates to an election and instances as voters:

- **Borda count:** solvers are ordered by CPUTIME and to each position is associated a score.
- **Range voting:** similar to Borda count, but using multiplicative positional weights.
- **Schulze's method:** it is a Condorcet method that computes the Schwartz set to determine a winner. We use an extension of the single overall winner procedure, in order to make it capable of generating an overall ranking.

Agenda

- Preliminaries
- Procedures
- YASM
- Comparative measures
- Conclusions

Yet Another Scoring Method

- YASMv2, improvement of YASM that combines:
  - traditional approach of the procedures used in automated reasoning systems contests
  - some ideas borrowed from voting systems

Yet Another Scoring Method

- YASMv2, improvement of YASM that combines:
  - traditional approach of the procedures used in automated reasoning systems contests
  - some ideas borrowed from voting systems
- Score $S_{s,i} = k_{s,i} \cdot (1 + H_i) \cdot \frac{L - T_{s,i}}{L - M_i}$

Yet Another Scoring Method

- YASMv2, improvement of YASM that combines:
  - traditional approach of the procedures used in automated reasoning systems contests
  - some ideas borrowed from voting systems
- Score $S_{s,i} = k_{s,i} \cdot (1 + H_i) \cdot \frac{L - T_{s,i}}{L - M_i}$
  - $k_{s,i}$: Borda-like positional weight
  - $(1 + H_i)$: relative hardness of the instance; it rewards the solvers that solve hard instances
  - $\frac{L - T_{s,i}}{L - M_i}$: relative speed of the solver with respect to the fastest solver on the instance; it rewards the solvers that are faster than other competitors

Yet Another Scoring Method

- YASMv2, improvement of YASM that combines:
  - traditional approach of the procedures used in automated reasoning systems contests
  - some ideas borrowed from voting systems
- Score $S_{s,i} = k_{s,i} \cdot (1 + H_i) \cdot \frac{L - T_{s,i}}{L - M_i}$
  - $k_{s,i}$: Borda-like positional weight
  - $(1 + H_i)$: relative hardness of the instance; it rewards the solvers that solve hard instances
  - $\frac{L - T_{s,i}}{L - M_i}$: relative speed of the solver with respect to the fastest solver on the instance; it rewards the solvers that are faster than other competitors
- Total score $S_s = \sum_i S_{s,i}$

Agenda

- Preliminaries
- Procedures
- YASM
- Comparative measures
- Conclusions

Homogeneity

- Degree of **(dis)agreement** between different aggregation procedures.

Homogeneity

- Degree of **(dis)agreement** between different aggregation procedures.
- Verify that the aggregation procedures considered
  - do not produce exactly the same solver rankings
  - do not yield antithetic solver rankings

Homogeneity

- Degree of **(dis)agreement** between different aggregation procedures.
- Verify that the aggregation procedures considered
  - do not produce exactly the same solver rankings
  - do not yield antithetic solver rankings
- Kendall rank correlation coefficient τ as measure of homogeneity.

Homogeneity

	CASC	QBF	SAT	YASM	YASMv2	Borda	r.v.	Schulze
CASC	–	1	0.71	0.86	0.79	0.86	0.71	0.86
QBF		–	0.71	0.86	0.79	0.86	0.71	0.86
SAT			–	0.86	0.86	0.71	0.71	0.71
YASM				–	0.86	0.71	0.71	0.71
YASMv2					–	0.86	0.86	0.86
Borda						–	0.86	1
r. v.							–	0.86
Schulze								–

r.v. = range voting

Fidelity

- Given a synthesized set of raw data, evaluates whether an aggregation procedure distorts the results.

Fidelity

- Given a synthesized set of raw data, evaluates whether an aggregation procedure distorts the results.
- Several samples of table RUNS filled with random results:
  - RESULT is assigned to SAT/UNSAT, TIME or FAIL with equal probability
  - a value of CPUTIME is chosen uniformly at random in the interval [0;1]

Fidelity

- Given a synthesized set of raw data, evaluates whether an aggregation procedure distorts the results.
- Several samples of table RUNS filled with random results:
  - RESULT is assigned to SAT/UNSAT, TIME or FAIL with equal probability
  - a value of CPUTIME is chosen uniformly at random in the interval [0;1]
- A high-fidelity aggregation procedure:
  - computes approximately the same scores for each solver
  - produces a final ranking where scores have a small variance-to-mean ratio

Fidelity

Method	Mean	Std	Median	Min	Max	IQ Range	F
QBF	182.25	7.53	183	170	192	13	88.54
CASC	182.25	7.53	183	170	192	13	88.54
SAT	87250	12520.2	83262.33	78532.74	119780.48	4263.94	65.56
YASM	46.64	2.22	46.33	43.56	51.02	2.82	85.38
YASMv2	1257.29	45.39	1268.73	1198.43	1312.72	95.11	91.29
Borda	984.5	127.39	982.5	752	1176	194.5	63.95
r. v.	12010.25	5183.86	12104	5186	21504	8096	24.12
SCHULZE	–	–	–	–	–	–	–

r.v. = range voting

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_1
INSTANCE_2
INSTANCE_3
INSTANCE_4
INSTANCE_5
INSTANCE_6
INSTANCE_7
INSTANCE_8
INSTANCE_9
INSTANCE_10
INSTANCE_11
INSTANCE_12
INSTANCE_13
INSTANCE_14
INSTANCE_15

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_1

INSTANCE_3

INSTANCE_6
INSTANCE_7
INSTANCE_8
INSTANCE_9

INSTANCE_11
INSTANCE_12

INSTANCE_14
INSTANCE_15

RDT-stability

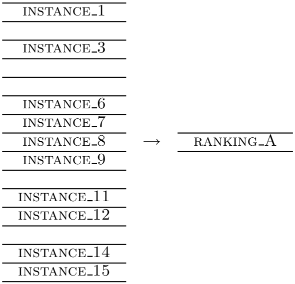
- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_1
INSTANCE_3
INSTANCE_6
INSTANCE_7
INSTANCE_8
INSTANCE_9
INSTANCE_11
INSTANCE_12
INSTANCE_14
INSTANCE_15

→

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.



RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_1
INSTANCE_2
INSTANCE_3
INSTANCE_4

INSTANCE_6
INSTANCE_7
→ RANKING_A

INSTANCE_10
INSTANCE_11
INSTANCE_12
INSTANCE_13

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_1
INSTANCE_2
INSTANCE_3
INSTANCE_4
INSTANCE_6
INSTANCE_7
INSTANCE_10
INSTANCE_11
INSTANCE_12
INSTANCE_13

→

RANKING_A
RANKING_B

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_2
INSTANCE_3
INSTANCE_4
INSTANCE_5
INSTANCE_6
INSTANCE_8
INSTANCE_9
INSTANCE_10
INSTANCE_14
INSTANCE_15

→

RANKING_A
RANKING_B

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

Test set		Rankings
INSTANCE_2		
INSTANCE_3		
INSTANCE_4		
INSTANCE_5		
INSTANCE_6		
		RANKING_A
INSTANCE_8	→	RANKING_B
INSTANCE_9		RANKING_C
INSTANCE_10		
INSTANCE_14		
INSTANCE_15		

RDT-stability

- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

INSTANCE_2
INSTANCE_3
INSTANCE_4
INSTANCE_5
INSTANCE_6
INSTANCE_8
INSTANCE_9
INSTANCE_10
INSTANCE_14
INSTANCE_15

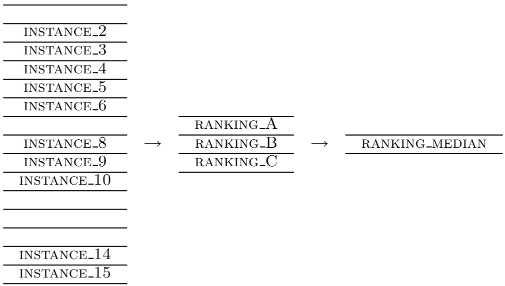
→

RANKING_A
RANKING_B
RANKING_C

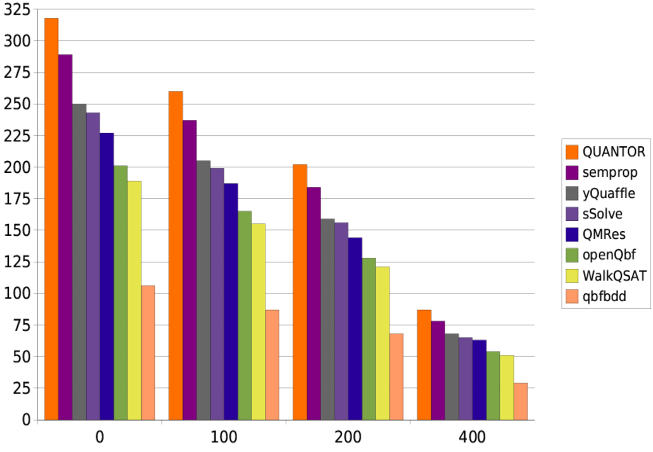
→

RDT-stability

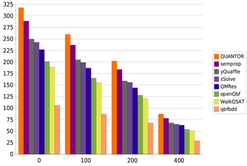
- Stability on a **Randomized Decreasing Test set** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the size of the original test set.

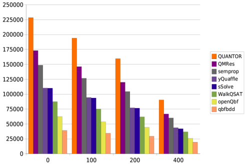


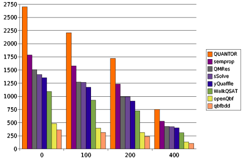
RDT-stability

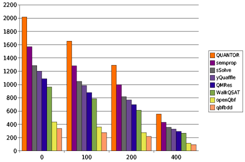


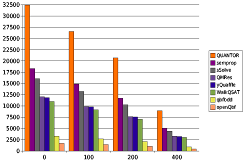
RDT-stability

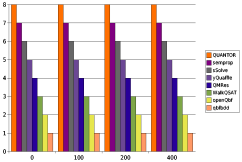










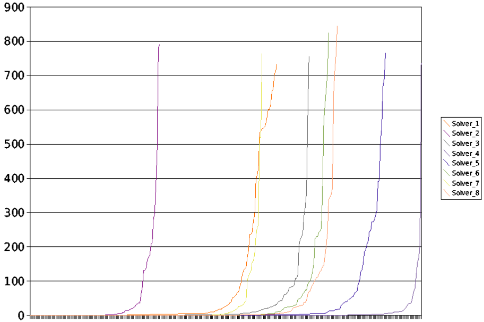


DTL-stability

- Stability on a **Decreasing Time Limit** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers.

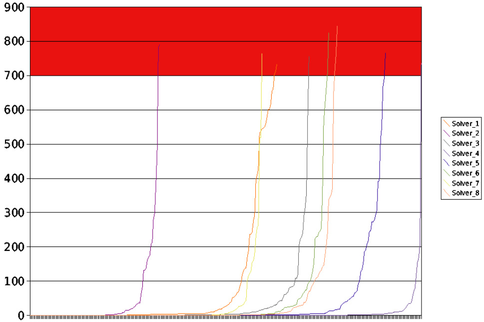
DTL-stability

- Stability on a **Decreasing Time Limit** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers.



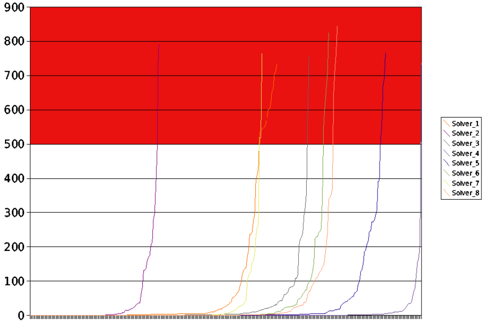
DTL-stability

- Stability on a **Decreasing Time Limit** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers.



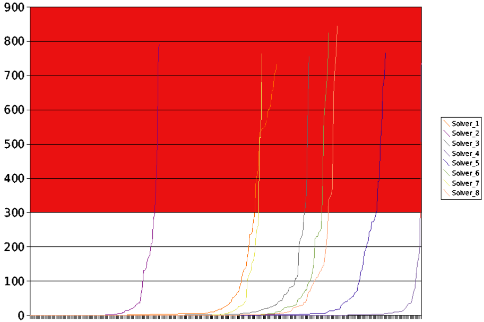
DTL-stability

- Stability on a **Decreasing Time Limit** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers.

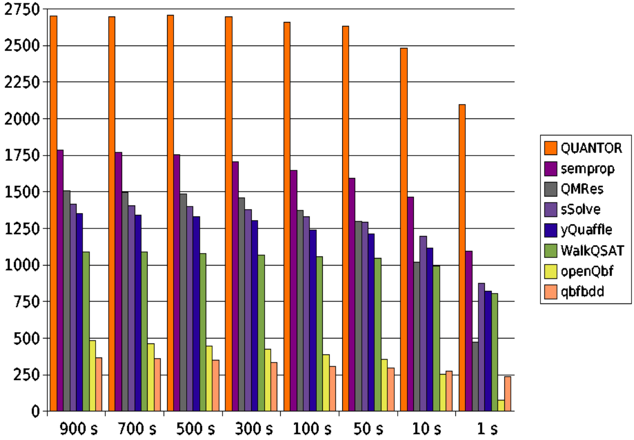


DTL-stability

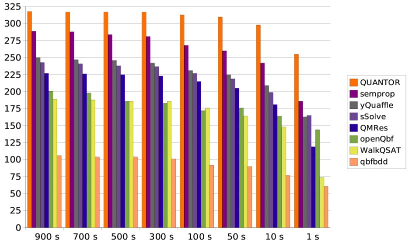
- Stability on a **Decreasing Time Limit** aims to measure how much an aggregation procedure is sensitive to perturbations that diminish the maximum amount of CPU time granted to the solvers.

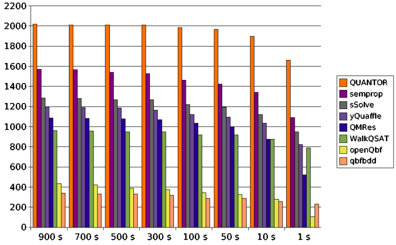


DTL-stability



DTL-stability





SBT-stability

- Stability on a **Solver Biased Test set** aims to measure how much an aggregation procedure is sensitive to a test set that is biased in favor of a given solver.

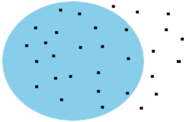
SBT-stability

- Stability on a **Solver Biased Test set** aims to measure how much an aggregation procedure is sensitive to a test set that is biased in favor of a given solver.

- Test set instances
- Solved by SOLVER_1
- Solved by SOLVER_2
- Solved by SOLVER_3

SBT-stability

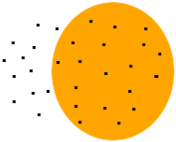
- Stability on a **Solver Biased Test set** aims to measure how much an aggregation procedure is sensitive to a test set that is biased in favor of a given solver.



- Test set instances
- Solved by SOLVER_1
- Solved by SOLVER_2
- Solved by SOLVER_3

SBT-stability

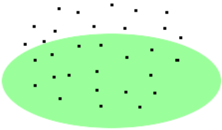
- Stability on a **Solver Biased Test set** aims to measure how much an aggregation procedure is sensitive to a test set that is biased in favor of a given solver.



- Test set instances
- Solved by SOLVER_1
- Solved by SOLVER_2
- Solved by SOLVER_3

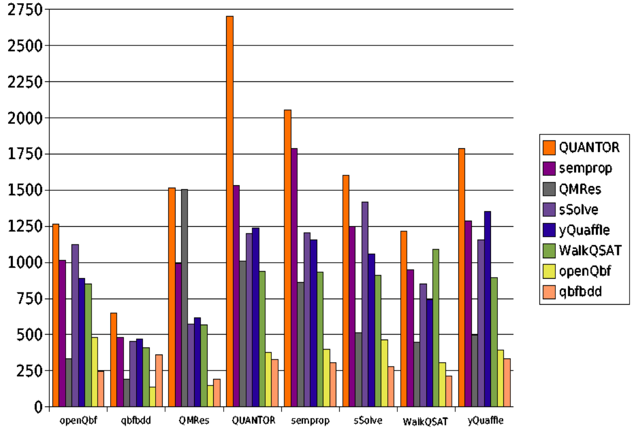
SBT-stability

- Stability on a **Solver Biased Test set** aims to measure how much an aggregation procedure is sensitive to a test set that is biased in favor of a given solver.



- Test set instances
- Solved by SOLVER_1
- Solved by SOLVER_2
- Solved by SOLVER_3

SBT-stability



SBT-stability

	CASC/QBF	SAT	YASM	YASMv2	Borda	r. v.	Schulze
OPENQBF	0.43	0.57	0.36	0.64	0.79	0.79	0.79
QBFBDD	0.43	0.43	0.36	0.64	0.79	0.86	0.79
QMRES	0.64	0.86	0.76	0.79	0.71	0.86	0.79
QUANTOR	1	0.86	0.86	0.86	0.93	0.86	0.93
SEMPROP	0.93	0.71	0.71	0.79	0.93	0.86	0.93
SSOLVE	0.71	0.57	0.57	0.79	0.86	0.79	0.86
WALKQSAT	0.57	0.57	0.43	0.71	0.64	0.79	0.79
YQUAFFLE	0.71	0.64	0.57	0.71	0.86	0.86	0.93
Mean	0.68	0.65	0.58	0.74	0.81	0.83	0.85

Kendall rank correlation coefficient between the test sets.

SOTA-relevance

- Relationship between the ranking obtained with an aggregation procedure and its SOTA-distance w.r.t. the SOTA solver.

SOTA-relevance

- Relationship between the ranking obtained with an aggregation procedure and its SOTA-distance w.r.t. the SOTA solver.
- The SOTA solver is the ideal solver that fares the best time for each instance among all solvers.

SOTA-relevance

- Relationship between the ranking obtained with an aggregation procedure and its SOTA-distance w.r.t. the SOTA solver.
- The SOTA solver is the ideal solver that fares the best time for each instance among all solvers.
- The SOTA-distance is the distance metric obtained by computing the Euclidean norm between the CPU times of any given solver and the SOTA solver.

SOTA-relevance

- Relationship between the ranking obtained with an aggregation procedure and its SOTA-distance w.r.t. the SOTA solver.
- The SOTA solver is the ideal solver that fares the best time for each instance among all solvers.
- The SOTA-distance is the distance metric obtained by computing the Euclidean norm between the CPU times of any given solver and the SOTA solver.

	SOTA-distance
CASC	1
QBF	1
SAT	0.71
YASM	0.86
YASM v2	0.79
Borda	0.86
range voting	0.71
Schulze	0.86

Agenda

- Preliminaries
- Procedures
- YASM
- Comparative measures
- Conclusions

Conclusions

- A larger test set is not necessarily a better test set (**RDT-stability**).

Conclusions

- A larger test set is not necessarily a better test set (**RDT-stability**).
- Increasing the time limit is not necessary useful, unless you increase it substantially (**DTL-stability**).

Conclusions

- A larger test set is not necessarily a better test set (**RDT-stability**).
- Increasing the time limit is not necessary useful, unless you increase it substantially (**DTL-stability**).
- The composition of the evaluation test set may heavily influence the final ranking (**SBT-stability**).

Conclusions

- Addition of the fidelity measure and improvement of the definition of SOTA-relevance.

Conclusions

- Addition of the fidelity measure and improvement of the definition of SOTA-relevance.
- YASMv2 is more powerful than YASM in terms of SBT-stability and fidelity.

Conclusions

- Addition of the fidelity measure and improvement of the definition of SOTA-relevance.
- YASMv2 is more powerful than YASM in terms of SBT-stability and fidelity.
- The fidelity measure shows the effectiveness of a hybrid approach such as YASMv2.

Possible Extensions

- Investigation in the explanatory power of the SOTA-distance metric.

Possible Extensions

- Investigation in the explanatory power of the SOTA-distance metric.
- Extension of the analysis to other aggregation procedures and/or voting systems.

Possible Extensions

- Investigation in the explanatory power of the SOTA-distance metric.
- Extension of the analysis to other aggregation procedures and/or voting systems.
- Investigation in the YASMv2 properties according to the framework of social choice theory.