

Active Learning for SAT Solver Benchmarking

Tobias Fuchs, Jakob Bach and Markus Iser

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.

*Corresponding author(s). E-mail(s): tobias.fuchs@kit.edu;
Contributing authors: jakob.bach@kit.edu; markus.iser@kit.edu;

Abstract

Benchmarking is a crucial phase in the development of algorithms. This also applies to SAT solvers for the problem of propositional satisfiability. Benchmark selection is about choosing representative problem instances that reliably discriminate solvers based on their runtime. In this paper, we present a dynamic benchmark selection approach based on active learning. Our approach predicts the rank of a new solver among its competitors with minimum runtime and maximum rank prediction accuracy. This approach was evaluated on the Anniversary track dataset from the 2022 SAT Competition. Our selection approach can predict the rank of a new solver after approximately 10 % of the time it would take to run the solver on all instances of this dataset, with a prediction accuracy of approximately 92 %. Additionally, the importance of instance families in the selection process is discussed. In conclusion, our tool offers a reliable method for solver engineers to assess a new solver’s performance in an efficient manner.

Keywords: Propositional satisfiability, Benchmarking, Active learning

1 Introduction

One of the main phases of algorithm engineering is benchmarking. This also applies to propositional satisfiability (SAT), the canonical \mathcal{NP} -complete problem. Benchmarking is, however, quite expensive regarding the runtime of experiments. While it is still feasible to benchmark a single or fixed amount of SAT solvers, developing new, competitive SAT solvers requires extensive experimentation with a variety of ideas, which often results in a combinatorial explosion of the configuration space of each algorithm [16]. In particular, a new solver idea is rarely best on the first try. Thus, it is highly desirable to reduce benchmarking time and discard unpromising ideas early,

allowing to test more approaches or spend more time on promising ones. The field of SAT solver benchmarking is well established, yet traditional benchmark selection approaches do not optimize benchmark runtime. Instead, they focus on selecting a representative set of instances for ranking solvers [11, 14]. With regard to the ranking of solvers, SAT Competitions typically employ the PAR-2 score, which, given a time limit of τ , is the average runtime with a penalty of 2τ for timeouts [8].

In this paper, we present a novel benchmark selection approach based on active learning. Our approach can predict the rank of a new solver with high accuracy in only a fraction of the time needed to evaluate the complete benchmark. Definition 1 specifies the problem we address. Note that our scenario assumes knowing the runtimes of all solvers, except the new one, on all instances. One could also imagine a collaborative filtering scenario, where runtimes are only partially known [22, 24].

Definition 1 (New-Solver Problem). *Let us consider a set of solvers \mathcal{A} , instances \mathcal{I} , and runtimes $r: \mathcal{A} \times \mathcal{I} \rightarrow [0, \tau]$ with each solver having a runtime limit of τ . For a new solver $\hat{a} \notin \mathcal{A}$, incrementally select benchmark instances from \mathcal{I} to maximize the confidence in predicting the rank of \hat{a} while minimizing the total benchmark runtime.*

The approach presented in this paper meets several criteria for benchmarking that are considered desirable. Rather than outputting a binary classification, namely whether the new solver is worse than an existing solver or not, we provide a scoring function that shows by which margin a solver is worse and how similar it is to existing solvers. In particular, our approach enables ranking the new solver amidst a set of existing solvers. We show that for ranking, it is not necessary to predict the exact per-instance runtime of the solvers, which would be a more challenging task. Furthermore, we optimize the *runtime* required for our strategy to arrive at its conclusion. To this end, we use instance *features* and known solver *runtimes*. Moreover, we select instances *non-randomly* and *incrementally*. In particular, we consider runtime information from already done experiments when choosing the next. By doing so, we can control the properties of the benchmarking approach, such as its required runtime. Our approach is *scalable* in that it ranks a new solver \hat{a} among any number of known solvers \mathcal{A} . In particular, we only subsample the benchmark once instead of comparing pairwise against each other solver [21].

We evaluated our approach using the SAT Competition 2022 Anniversary Track dataset [2], consisting of 5355 instances and runtimes of 28 solvers. Cross-validation was performed by treating each solver once as the new solver and learning to predict the PAR-2 rank of that solver. On average, our predictions achieve an accuracy of approximately 92% with only approximately 10% of the runtime required to evaluate these solvers on the complete set of instances. Our entire source code¹ and experimental data² are available on GitHub.

This article is an extended and refined version of our conference paper with the same title, “Active Learning for SAT Solver Benchmarking” [9]. In particular, we significantly extended the analysis of runtime-prediction approaches, which

¹<https://github.com/mathefuchs/al-for-sat-solver-benchmarking>

²<https://github.com/mathefuchs/al-for-sat-solver-benchmarking-data>

led to our choice of a solver runtime model for use in benchmark instance selection (cf. Section 4.4). Furthermore, we expanded the evaluation of which benchmark instances are selected by our approach (cf. Section 5.3).

2 Related Work

Benchmarking is of great interest in numerous research areas and represents an active field of research in its own right. Studies have shown that the compilation of benchmark instances poses a number of challenges. The use of biased benchmarks can easily lead to fallacious interpretations [7]. Benchmarking also involves a number of interchangeable elements, including the performance measures used, the way the measures are aggregated, and the way missing values are handled. It is possible that questionable research practices could alter these elements a-posteriori in order to meet expectations, thereby skewing the results [26].

The following sections present a discussion of related work from the areas of static benchmark selection, algorithm configuration, incremental benchmark selection, and active learning. Table 1 compares the most relevant approaches, which all pursue slightly different goals. Thus, our approach is *not* a general improvement over the others but the only one fully aligned with Definition 1.

2.1 Static Benchmark Selection

Benchmark selection is essential for competitions such as the SAT Competition. In such competitions, the organizers define the rules for composing the benchmarks. These selection strategies are primarily static, in that they do not depend on particular solvers to distinguish. Balint et al. provide an overview of benchmark-selection criteria in different solver competitions [1]. Froleys et al. describe benchmark selection in recent SAT competitions [8]. Manthey and Möhle propose a feature-based approach to remove redundancy from competition benchmarks [20], while Mısırlı presents a similar feature-based approach to reduce benchmarks by matrix factorization and clustering [23].

In their study, Hoos et al. [14] discuss the desirable properties of SAT benchmark instances. They identify three key selection criteria: instance variety to avoid over-fitting, adapted instance hardness (not too easy but also not too hard), and the avoidance of duplicate instances. To filter instances that are too similar, they employ a distance-based approach with the SATzilla features [36, 37]. It should be noted, however, that the approach does not optimize for benchmark runtime. Instead, instances are selected at random, with the exception of constraints on instance hardness and feature distance.

2.2 Algorithm Configuration

Further related work can be found within the field of algorithm configuration [15, 31], as exemplified by the configuration system SMAC [16]. Thereby, the goal is to tune SAT solvers for a given sub-domain of problem instances. While this task differs from our goal, for instance, in that we do not need to navigate the configuration space,

Table 1: Comparison of features of our benchmark-selection approach, the static benchmark-selection approach by Hoos et al. [14], the algorithm configuration system SMAC [16], and the active-learning approaches by Matricon et al. [21].

Feature	Hoos [14]	SMAC [16]	Matricon [21]	Our approach
Ranking/Scoring	✓	✗	(✓)	✓
Runtime Minimization	✗	✓	✓	✓
Incremental/Non-Random	✗	✗	✓	✓
Scalability	✓	✓	✗	✓

there are similarities to our approach as well. For example, SMAC employs an iterative, model-based selection procedure similar to our approach, but this is for selecting configurations rather than instances. However, an algorithm configurator cannot be used to *rank* or *score* a new solver since algorithm configuration is designed to identify the optimal configuration. Furthermore, while a model-based selection strategy is employed to sample configurations, instance selection is made *randomly*, i.e., without building a model over instances.

2.3 Incremental Benchmark Selection

Matricon et al. present an incremental benchmark selection approach [21]. Their *per-set efficient algorithm selection problem* (PSEAS) is similar to our *New-Solver Problem*, as given in Definition 1. To differentiate between a *pair of* SAT solvers, the authors propose an iterative approach to select a solver-specific subset of instances until a desired confidence level is reached. This is achieved by calculating a scoring metric for all unselected instances, running the experiment with the highest score, and updating the confidence. Their approach ticks off most of the desired features in Table 1. However, the approach only compares solvers binarily rather than providing a *scoring*. Consequently, it is unclear how similar two given solvers are or on which instances they behave similarly. Furthermore, a significant shortcoming is the lack of *scalability* with the number of solvers. Comparing only pairs of solvers, evaluating a new solver requires sampling a separate benchmark for each existing solver. In contrast, our approach allows comparing a new solver against a set of existing solvers by sampling only one benchmark. In contrast to our approach, which allows for the comparison of a new solver against a set of existing solvers by sampling only one benchmark, their pairwise method requires the evaluation of a new solver to be conducted on a separate benchmark for each existing solver.

2.4 Active Learning

In passive machine learning, prediction models are trained on datasets with given instance labels (cf. Fig. 1a). In contrast, active learning (AL) begins with no or little labeled data. AL then repeatedly selects interesting problem instances for which to acquire labels, gradually improving the prediction model (cf. Fig. 1b). AL methods are particularly advantageous when acquiring labels is computationally expensive, such as obtaining solver runtimes. Without AL methods, it is not evident which instances to label and which not. The objective is to maximize the utility of instances to our

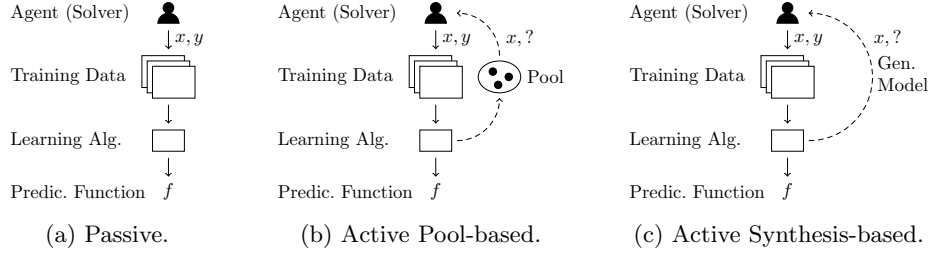


Fig. 1: Types of machine learning (depiction inspired by Rubens et.al. [28]).

model, which is measured by the rank prediction accuracy. Concurrently, the objective is to minimize the cost of label acquisition, which is measured by the instance’s predicted runtime. Consequently, the objective is to develop an accurate prediction model without the necessity of labeling each data point.

Rubens et. al. [28] survey active-learning advances. While synthesis-based AL methods [5, 10, 33] generate instances for labeling, pool-based methods [12, 13, 19] rely on a fixed set of unlabeled instances to sample from. Recent synthesis-based methods within the field of SAT solving demonstrate the generation of problem instances with desired properties [5, 10]. However, this objective is distinct from ours. While those approaches aim to generate instances on which a solver is either effective or ineffective, our objective is to predict whether a solver is effective or ineffective on an *existing* benchmark. In their work, Volpato and Guangyan employ pool-based AL to train an instance-specific algorithm selector [34]. Rather than benchmarking a solver’s overall performance, their objective is to recommend the best solver for each SAT instance from a set of available solvers.

3 Active Learning for SAT Solver Benchmarking

Algorithm 1 outlines our benchmarking framework. Given a set of solvers \mathcal{A} , instances \mathcal{I} and runtimes r , we first initialize a prediction model \mathcal{M} for the new solver $\hat{a} \notin \mathcal{A}$ (Line 1). The prediction model \mathcal{M} is used to select instances repeatedly (Line 4) for benchmarking the new solver \hat{a} (Line 5). The acquired result is subsequently employed to update the prediction model, denoted by \mathcal{M} (Line 7). When the stopping criterion is met (Line 3), the benchmarking loop is terminated and the final score of \hat{a} is estimated (Line 8). This score, along with the acquired instances and runtime measurements, is returned by Algorithm 1.

Section 3.1 provides an overview of the underlying prediction model \mathcal{M} and outlines the methodology for deriving a solver ranking from it. Section 3.2 presents the criteria for selecting instances. Finally, Section 3.3 presents potential stopping conditions.

3.1 Solver Model

The model \mathcal{M} provides a runtime-label prediction function $f : \hat{\mathcal{A}} \times \mathcal{I} \rightarrow \mathbb{R}$ for all solvers $\hat{\mathcal{A}} := \mathcal{A} \cup \{\hat{a}\}$. This prediction function powers instance selection as described

Algorithm 1: Incremental Benchmarking Framework

Input: Solvers \mathcal{A} , Instances \mathcal{I} , Runtimes $r : \mathcal{A} \times \mathcal{I} \rightarrow [0, \tau]$, Solver \hat{a}

Output: Predicted Score of \hat{a} , Measured Runtimes \mathcal{R}

```
1  $\mathcal{M} \leftarrow \text{initModel}(\mathcal{A}, \mathcal{I}, r, \hat{a})$  // cf. Section 3.1
2  $\mathcal{R} \leftarrow \emptyset$ 
3 while not stop( $\mathcal{M}$ ) do // cf. Section 3.3
4    $e \leftarrow \text{selectNextInstance}(\mathcal{M})$  // cf. Section 3.2
5    $t \leftarrow \text{runExperiment}(\hat{a}, e)$  // Runs  $\hat{a}$  on  $e$  with timeout  $\tau$ 
6    $\mathcal{R} \leftarrow \mathcal{R} \cup \{(e, t)\}$ 
7    $\text{updateModel}(\mathcal{M}, \mathcal{R})$  // cf. Section 3.1
8  $s_{\hat{a}} \leftarrow \text{predictScore}(\mathcal{M})$  // cf. Section 3.1
9 return ( $s_{\hat{a}}, \mathcal{R}$ )
```

in Section 3.2. During model updates (Algorithm 1, Line 7), the prediction function f is trained to predict a transformed version of the acquired runtime data set \mathcal{R} . The transformation of the runtime data set is described in the subsequent section. The features described in Section 4.2 serve as the input to the model. Furthermore, it should be noted that a new prediction model is constructed in each iteration, as the runtime of the experiments (Line 5) dominates the training time of the model by a significant margin. Finally, the score of the new solver, denoted by \hat{a} , is predicted with the prediction function f (Line 8).

3.1.1 Runtime Transformation

For the prediction model M , we transform the real-valued runtimes into discrete runtime labels on a per-instance basis. For each instance $e \in \mathcal{I}$, we use a clustering algorithm to assign the runtimes in $\{r(a, e) \mid a \in \mathcal{A}\}$ to one of k clusters C_1, \dots, C_k such that the fastest runtimes for the instance e are in cluster C_1 and the slowest are in cluster C_{k-1} . Timeouts τ always form a separate cluster C_k . The runtime transformation function $\gamma_k : \mathcal{A} \times \mathcal{I} \rightarrow \{1, \dots, k\}$ is then specified as follows:

$$\gamma_k(a, e) = j \Leftrightarrow r(a, e) \in C_j$$

In the context of such a clustering, for an instance $e \in \mathcal{I}$, a solver $a \in \mathcal{A}$ is defined as belonging to the $\gamma_k(a, e)$ -fastest solvers on instance e . Empirical studies on portfolio solvers have demonstrated that discretization is an effective approach in practice [4, 25]. Section 4.4 presents a detailed analysis of the runtime transformation we designed for our approach. The analysis demonstrates that our approach achieves higher accuracy for discrete runtime labels than for raw runtimes.

3.1.2 Ranking Solvers

To determine the rank of solvers, we apply PAR-2 scoring using the discrete runtime labels $\gamma_k(a, e)$. The adapted scoring function $s_k : \mathcal{A} \rightarrow [1, 2 \cdot k]$ induces a ranking

among solvers and is depicted in Equation (1).

$$s_k(a) := \frac{1}{|\mathcal{I}|} \sum_{e \in \mathcal{I}} \gamma'_k(a, e) \quad \gamma'_k(a, e) := \begin{cases} 2 \cdot \gamma_k(a, e) & \text{if } \gamma_k(a, e) = k \\ \gamma_k(a, e) & \text{otherwise} \end{cases} \quad (1)$$

3.2 Instance Selection

Selecting an instance based on the model is a core functionality of our framework (cf. Algorithm 1, Line 4). In this section, we introduce two instance sampling strategies, one that minimizes uncertainty and one that maximizes information gain. Both strategies use the model's label-prediction function f and are inspired by existing work within the realms of active learning [29]. These methods require the model's predictions to include probabilities for the k discrete runtime labels. Let $f' : \hat{\mathcal{A}} \times \mathcal{I} \rightarrow [0, 1]^k$ denote this modified prediction function. In the following, the set $\tilde{\mathcal{I}} \subseteq \mathcal{I}$ denotes the instances that have already been sampled.

Uncertainty Sampling

The uncertainty sampling strategy selects the instance that is closest to the model's decision boundary. This is achieved by selecting the instance $e \in \mathcal{I} \setminus \tilde{\mathcal{I}}$ that minimizes $U(e)$, which is specified by the following equation.

$$U(e) := \left| \frac{1}{k} - \max_{n \in \{1, \dots, k\}} f'(\hat{a}, e)_n \right|$$

Information-Gain Sampling

The information-gain sampling strategy selects the instance with the highest expected entropy reduction regarding the runtime labels of the instance. To be more precise, we select the instance $e \in \mathcal{I} \setminus \tilde{\mathcal{I}}$ that maximizes $IG(e)$, which is specified in the following equation.

$$IG(e) := H(e) - \sum_{n=1}^k f'(\hat{a}, e)_n \hat{H}_n(e)$$

In the equation, $H(e)$ denotes the entropy of the runtime labels $\gamma(a, e)$ over all $a \in \mathcal{A}$ and $H(e, n)$ denotes the entropy of these labels plus n as the runtime label for \hat{a} . The term $\hat{H}_n(e)$ is computed for every possible runtime label $n \in \{1, \dots, k\}$. By maximizing the information gain, instances are selected that identify solvers with similar behavior.

3.3 Stopping Criteria

In this section, we present the two dynamic stopping criteria employed in our experiments: the Wilcoxon and the ranking stopping criterion (cf. Algorithm 1, Line 3).

Wilcoxon Stopping Criterion

The Wilcoxon stopping criterion is employed to halt the active-learning process based on the confidence that the predicted runtime labels of the new solver are sufficiently

different from those of existing solvers. This criterion is loosely inspired by Matricon et. al. [21]. The average p -value $W_{\hat{a}}$ of a Wilcoxon signed-rank test $w(S, P)$, as given by the following equation, of the two runtime label distributions $S = \{\gamma(a, e) \mid e \in \mathcal{I}\}$ for an existing solver a and $P = \{f(\hat{a}, e) \mid e \in \mathcal{I}\}$ of the new solver \hat{a} is used to assess the statistical significance of the predicted runtime labels.

$$W_{\hat{a}} := \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} w(S, P)$$

In order to improve the stability of this criterion, an exponential moving average is employed to mitigate the impact of outliers, with the process terminated when the value of $W_{\text{exp}}^{(i)}$ drops below a fixed threshold.

$$\begin{aligned} W_{\text{exp}}^{(0)} &:= 1 \\ W_{\text{exp}}^{(i)} &:= \beta W_{\hat{a}} + (1 - \beta) W_{\text{exp}}^{(i-1)} \end{aligned}$$

Ranking Stopping Criterion

In contrast, the stopping criterion for the ranking process is less sophisticated than the previous one. In this case, the active learning process is terminated if the ranking induced by the model’s predictions (Equation 1) remains unchanged within the last l iterations. However, it should be noted that the concrete values of the predicted score $s_{\hat{a}}$ may still change. Nevertheless, in this case, the induced ranking is of sole interest.

4 Experimental Design

In light of the preceding instantiations of Algorithm 1, this section outlines the experimental design employed in the presented study. This includes an overview of the evaluation framework, the data sets employed, the hyperparameter selections, and the implementation details.

4.1 Evaluation Framework

As stated in the Introduction, this work addresses the *New-Solver Problem* (cf. Definition 1). As described in Section 3.1, a prediction model provides us with an estimated scoring value for the new solver \hat{a} . To evaluate the efficacy of a concrete instantiation of Algorithm 1, which entails a specific choice for all subroutines, we perform cross-validation on our set of solvers.

This is demonstrated in Algorithm 2. Each solver assumes the role of the new solver \hat{a} once (Line 2) and is consequently excluded from the set of solvers \mathcal{A} (Line 3). After running our active-learning framework for solver \hat{a} (Line 4), we compute the value of both our optimization goals, i.e., ranking accuracy and runtime. The *ranking accuracy*, expressed in the interval $[0, 1]$ (higher is better), is defined as the fraction of correctly ranked pairs of elements, namely, (\hat{a}, a) , where a belongs to the set of all solvers \mathcal{A} ranked by the ground-truth scoring function par_2 (Lines 5-8). The *fraction of*

Algorithm 2: Evaluation Framework

Input: Solvers \mathcal{A} , Instances \mathcal{I} , Runtimes $r : \mathcal{A} \times \mathcal{I} \rightarrow [0, \tau]$
Output: Average Ranking Accuracy \bar{O}_{acc} , Average Fraction of Runtime \bar{O}_{rt}

```
1  $O \leftarrow \emptyset$ 
2 for  $\hat{a} \in \mathcal{A}$  do
3    $\mathcal{A}' \leftarrow \mathcal{A} \setminus \{\hat{a}\}$ 
4    $(s_{\hat{a}}, \mathcal{R}) \leftarrow \text{runALAlgorithm}(\mathcal{A}', \mathcal{I}, r, \hat{a})$  // Refer to Algorithm 1
   // Determine Ranking Accuracy
5    $O_{\text{acc}} \leftarrow 0$ 
6   for  $a \in \mathcal{A}$  do
7     if  $(s_k(a) - s_{\hat{a}}) \cdot (\text{par}_2(a) - \text{par}_2(\hat{a})) > 0$  then
8        $O_{\text{acc}} \leftarrow O_{\text{acc}} + \frac{1}{|\mathcal{A}|}$ 
   // Determine Runtime Fraction
9    $r \leftarrow \sum_{e \in \mathcal{I}} r(\hat{a}, e)$ 
10   $O_{\text{rt}} \leftarrow 0$ 
11  for  $e \in \mathcal{I}$  do
12    if  $\exists t, (e, t) \in \mathcal{R}$  then
13       $O_{\text{rt}} \leftarrow O_{\text{rt}} + \frac{t}{r}$ 
14   $O \leftarrow O \cup \{(O_{\text{acc}}, O_{\text{rt}})\}$ 
15   $(\bar{O}_{\text{acc}}, \bar{O}_{\text{rt}}) \leftarrow \text{average}(O)$ 
16 return  $(\bar{O}_{\text{acc}}, \bar{O}_{\text{rt}})$ 
```

runtime required by the algorithm to reach its conclusion is represented by O_{rt} , which lies within the interval $[0, 1]$ (lower is better). This metric compares the accumulated runtimes over the sampled instances to the accumulated runtimes over all instances in the dataset (Lines 9-13). Once all cross-validation results have been collected (Line 14), the output metrics are averaged (Line 15).

The procedure is designed to identify an approach that maximizes Equation (2), in which $\delta \in [0, 1]$ is used for linear weighting, with weighting factor $\delta \in [0, 1]$, between the two optimization goals O_{acc} and O_{rt} . The plotting of the approaches that maximize O_{δ} for all $\delta \in [0, 1]$ on an O_{rt} - O_{acc} -diagram provides us with a Pareto front of the best approaches for different optimization-goal weightings.

$$O_{\delta} := \delta O_{\text{acc}} + (1 - \delta) (1 - O_{\text{rt}}) \quad , \quad (2)$$

4.2 Data

In our experiments we work with the dataset of the SAT Competition 2022 Anniversary Track [2]. The dataset consists of 5355 instances with corresponding runtime data of 28 sequential SAT solvers. We also use a database of 56 instance features³ from

³https://benchmark-database.de/getdatabase/base_db

the Global Benchmark Database (GBD) by Iser et al. [17]. They include instance size features and node distribution statistics for several graph representations of SAT instances, and are primarily inspired by the SATzilla 2012 features described in [37]. All features are numeric and have no missing values. We drop 10 of the 56 features due to zero variance. In total, the prediction models have access to 46 instance features and 27 out of 28 runtime features, excluding the respective new solver \hat{a} .

In addition, we retrieve instance family information⁴ to evaluate the composition of our sampled benchmarks. Instance families consist of instances from the same application domain, such as planning, cryptography, etc., and are a valuable tool for analyzing solver performance. For hyperparameter tuning, we randomly sample 10% of the complete set of 5355 instances with stratification regarding the instance family. All instance families that are too *small*, i.e., 10% of them correspond to less than one instance, are put into a meta-family for stratification. This *tuning dataset* allows a more extensive exploration of the hyperparameter space.

4.3 Hyperparameters

Given the Algorithm 1, there are several possible instantiations for the three subroutines *ranking*, *selection*, and *stopping*. These experimental configurations are described in the following.

Ranking. Regarding *ranking*, we experiment with the following instantiations.

- Observed PAR-2 ranking of already sampled instances
- Predicted runtime-label ranking
 - History size: Consider the latest 1, 10, 20, 30, or 40 predictions within a voting approach for stability. The latest x predictions for each instance vote on the instance’s winning label.
 - Fallback threshold: If the difference of scores between the new solver \hat{a} and another solver drops below 0.01, 0.05, or 0.1, use the partially observed PAR-2 ranking as a tie-breaker.

Selection. We experiment with the following instantiations of instance *selection*. We experiment with the following instantiations of the instance *selection* criteria. Since the potential runtime of the experiments is orders of magnitude larger than the model update time, we consider incrementing our benchmark one instance at a time rather than using batches, which is also proposed in recent advances in active learning [30, 33]. A drawback of this is the lack of parallel execution of runtime experiments.

- Random sampling
- Uncertainty sampling
 - Fallback threshold: Use random sampling for the first 0%, 5%, 10%, 15%, or 20% of instances to explore the instance space.
 - Runtime scaling: Whether to normalize uncertainty scores per instance by the average runtime of solvers on it or use the absolute values.
- Information-gain sampling

⁴https://benchmark-database.de/getdatabase/meta_db

- Fallback threshold: Use random sampling for the first 0 %, 5 %, 10 %, 15 %, or 20 % of instances to explore the instance space.
- Runtime scaling: Whether to normalize information-gain scores per instance by the average runtime of solvers on it or use the absolute values.

Stopping. We evaluate the following *stopping* criteria.

- Subset-size stopping criterion, using 10 % or 20 % of instances
- Ranking stopping criterion
 - Minimum amount: Sample at least 2 %, 8 %, 10 %, or 12 % of instances before applying the criterion.
 - Convergence duration: Stop if the predicted ranking stays the same for a number of sampled instances equal to 1 % or 2 % of all instances.
- Wilcoxon stopping criterion
 - Minimum amount: Sample at least 2 %, 8 %, 10 %, or 12 % of instances before applying the criterion.
 - Average of p -values to drop below: 5 %.
 - Exponential-moving average: Incorporate previous significance values by using an EMA with $\beta = 0.1$ or $\beta = 0.7$.

4.4 Runtime Discretization and Prediction

One of the most important parts of our framework is the choice of solver runtime model to inform active learning selection decisions. In this section, we detail how clustering solver runtimes can lead to nuanced runtime labels, and which machine learning models are best at predicting them.

As a baseline for improvement, Table 2 shows the runtime prediction performance of two regression models, a random forest regressor and an MLP regressor, using 46 instance features and 27 known solver runtimes. Both models were chosen for their robustness and ability to capture nonlinear dependencies [3], and both perform poorly, as evidenced by the magnitudes of the mean regression errors. An RMSE error of about 2000 s is quite poor compared to solver runtimes that range from 0 to 5000 s. Therefore, we decided to discretize the runtimes.

Table 3 shows the performance of different models in predicting timeouts. While the second column shows the performance of models trained on the 46 instance features, the third column shows the performance of models trained on the 46 instance features and the 27 runtime features. While the Quadratic DA model performs poorly on the 46 instance features, it performs well on the 46 instance features and the 27

Table 2: Regression performance of models trained on 46 instance and 27 runtime features. We report the RMSE and MAE metrics on the test sets.

Regression Models	Avg. RMSE (\pm std)	Avg. MAE (\pm std)
Random-Forest-Regressor	1929.84 s (\pm 1161.15 s)	957.36 s (\pm 297.21 s)
MLP-Regressor	2275.20 s (\pm 1545.01 s)	1284.30 s (\pm 390.57 s)

Table 3: MCC performance of models predicting timeouts. The models are trained 28 times with each solver as a target once.

Timeout Prediction Models	46 Instance Features Avg. MCC (\pm std)	46 + 27 Runtime Feat. Avg. MCC (\pm std)
Stacking (QDA + RF)	0.6513 (\pm 0.0458)	0.9527 (\pm 0.0292)
Quadratic-DA (QDA)	0.2593 (\pm 0.1310)	0.9290 (\pm 0.0339)
Random-Forest-Classifer (RF)	0.6607 (\pm 0.0547)	0.8530 (\pm 0.0479)
Ada-Boost	0.5412 (\pm 0.0612)	0.8384 (\pm 0.0444)
Decision-Tree	0.5980 (\pm 0.0423)	0.8059 (\pm 0.0707)
Logistic-Regression	0.2031 (\pm 0.0728)	0.8052 (\pm 0.1018)
k NN-Classifer	0.5108 (\pm 0.0387)	0.7885 (\pm 0.1521)
MLP-Classifer	0.1293 (\pm 0.0718)	0.7760 (\pm 0.1408)
Support-Vector-Machines	0.0595 (\pm 0.0554)	0.7757 (\pm 0.2149)
Naive-Bayes	0.1173 (\pm 0.0872)	0.7306 (\pm 0.1394)

runtime features. However, the best prediction performance is achieved by a stacking ensemble [35], which combines quadratic discriminant analysis [32] and a random forest classifier [3]. Stacking means that another prediction model, in our case a simple decision tree, decides which of the two ensemble members makes the prediction about which instance.

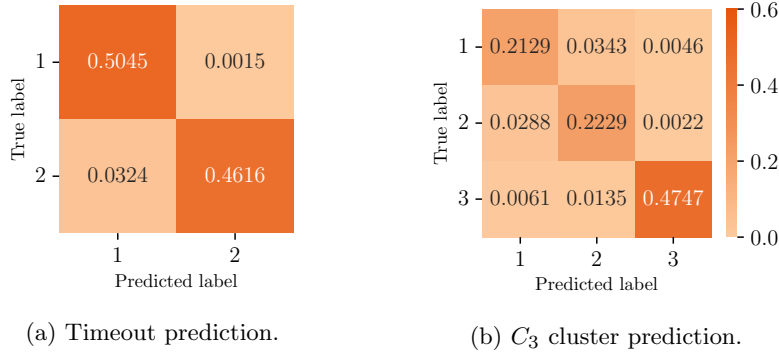
Since runtime prediction is quite challenging, we discretize the runtimes. To define prediction targets for each instance, we cluster the solver runtimes into k clusters. We denote this adapted solver scoring function which assigns a solver to its cluster for each instance with s_k as defined in Equation (1). In preliminary experiments, the most *useful* labels were produced with hierarchical clustering and a log-single-link criterion. In our chosen hierarchical procedure, each non-timeout runtime starts in a separate interval. We then gradually merge intervals with the smallest single-link logarithmic distance until the desired number of partitions is reached. The clustering is instance-specific, i.e., the bounds are chosen to produce groups of solvers with similar performance for each instance. Other clustering approaches we have tried include hierarchical clustering with mean, median, and full-link criteria, as well as k -means and spectral clustering.

We evaluated the performance of the best models from Table 3 in predicting the such determined s_k . Table 4 shows their performance in predicting the C_k clusters for $k \in \{3, 4, 5\}$. While the second column shows the prediction performance of the models trained on the 46 instance features including the 27 known solver runtimes, the third column shows the prediction performance of the models trained on these features plus 27 features representing the known C_k clustering labels of the known solver runtimes. We observe that the models perform better when the clustering labels are included as features. In addition, the prediction performance degrades rapidly as the number of clusters increases. In all cases, the stacking ensemble performs best.

Our final choice for the prediction model is the stacking ensemble of a quadratic discriminant analysis and a random forest. Both of these models can learn non-linear relationships between the instance features and the runtime labels. Figure 2 shows the

Table 4: MCC performance of the best models predicting the C_k clusterings.

C_3 Cluster Prediction Models	46 + 27 Runtime Feat. Avg. MCC (\pm std)	46 + 27 + 27 Clus. Feat. Avg. MCC (\pm std)
Stacking (QDA + RF)	0.7464 (\pm 0.0497)	0.8380 (\pm 0.0634)
Quadratic-DA (QDA)	0.6903 (\pm 0.0607)	0.7738 (\pm 0.0459)
Random-Forest-Classifer (RF)	0.7116 (\pm 0.0385)	0.7841 (\pm 0.0469)
C_4 Cluster Prediction Models	46 + 27 Runtime Feat. Avg. MCC (\pm std)	46 + 27 + 27 Clus. Feat. Avg. MCC (\pm std)
Stacking (QDA + RF)	0.6562 (\pm 0.0538)	0.7366 (\pm 0.0377)
Quadratic-DA (QDA)	0.6007 (\pm 0.0765)	0.6872 (\pm 0.0469)
Random-Forest-Classifer (RF)	0.6033 (\pm 0.0455)	0.6828 (\pm 0.0527)
C_5 Cluster Prediction Models	46 + 27 Runtime Feat. Avg. MCC (\pm std)	46 + 27 + 27 Clus. Feat. Avg. MCC (\pm std)
Stacking (QDA + RF)	0.5835 (\pm 0.0615)	0.6396 (\pm 0.0553)
Quadratic-DA (QDA)	0.5508 (\pm 0.0669)	0.5881 (\pm 0.0912)
Random-Forest-Classifer (RF)	0.5420 (\pm 0.0459)	0.5895 (\pm 0.0475)

**Fig. 2:** Confusion matrices for predicting timeouts and C_3 clusters.

confusion matrices of the stacking ensemble for predicting timeouts and C_3 clusters. Both perform reasonably well, with the C_3 cluster prediction model performing slightly worse than the timeout prediction model. Notably, for the C_3 prediction model, there is less confusion between the non-timeout labels (labels 1 and 2) and the timeout label (label 3) than between the non-timeout labels (labels 1 and 2). This suggests that deciding on timeouts is easier than distinguishing between different groups of solvers.

To ensure that we obtained *useful* labels, we evaluated whether the discretized labels still discriminate solvers and agree with the actual PAR-2 ranking. Table 5 shows the ranking of the solvers in the SAT Competition 2022 Anniversary Track [2] according to the standard PAR-2 runtime score and according to our discretized s_3 score. We observe that our ranking approach correctly decides almost all (about 97.45 %;

Table 5: Ranking of Anniversary-Track solvers with standard PAR-2 runtime and our discretized s_3 ranking.

PAR-2		s_3		Solver
Rank	Score	Rank	Score	
1	2808.13	1	1.1717	kissat-mab-esa
2	2812.93	2	1.1832	kissat-sc2022-bulky
3	2835.25	3	1.1862	ekissat-mab-gb-db
4	2835.59	4	1.1868	kissat-mab-ucb
5	2836.92	5	1.1868	kissat-inc
6	2845.19	6	1.1926	ekissat-mab-db-v1
7	2846.73	7	1.1930	kissat-mab-moss
8	2857.67	8	1.1947	kissat-mab-hywalk
9	2869.45	9	1.1998	kissat-sc2022-light
10	2899.70	10	1.2164	kissat-els-v2
11	2953.59	11	1.2290	hkis-unsat
12	2967.53	12	1.2347	kissat-adaptive-restart
13	2976.56	13	1.2475	seqfrost-noextend
14	3014.40	16	1.2645	kissat-els-v1
15	3017.73	14	1.2509	cadical-esa
16	3036.83	15	1.2613	cadical-reorder
17	3049.90	20	1.3648	cadical-rel-scavel
18	3080.66	19	1.2965	kissat-relaxed
19	3095.73	17	1.2815	cadical-dvdl-v1
20	3101.12	18	1.2856	cadical-dvdl-v2
21	3273.95	21	1.3786	glucose-reboot
22	3290.90	25	1.4707	lstech-maple-hywalk
23	3292.68	23	1.4478	lstech-maple
24	3400.72	24	1.4693	slime-sc-2022-beta
25	3412.11	26	1.5237	slime-sc-2022
26	3436.07	22	1.4161	hcad-v1-psids
27	3506.32	27	1.5433	maple-lcmdistchrbt-dl-v3
28	4741.50	28	2.0581	isasat

$\sigma = 3.68\%$) solver pairs, i.e. which solver is faster among two given solvers. In particular, the Spearman correlation of s_3 and PAR-2 ranking is about 0.988, which is very close to the optimal value of 1 [6].

We also analyzed how runtime discretization affects the distinguishability of solver pairs. Figure 3 shows the confusion matrices of solver pairs with significant differences in their PAR-2 scores, timeouts, and s_3 scores. According to a Wilcoxon signed-rank test with a level of $\alpha = 0.05$, 87.83% of the solver pairs have significantly different scores after discretization, only a small decrease compared to 89.95% before discretization. These results support the usefulness of discretized runtimes for our framework.

4.5 Implementation Details

For reproducibility, our source code and data are available on GitHub (see footnotes in Section 1). Our code is implemented in PYTHON using *scikit-learn* [27] for making predictions and *gbd-tools* [17] for retrieving SAT instances.

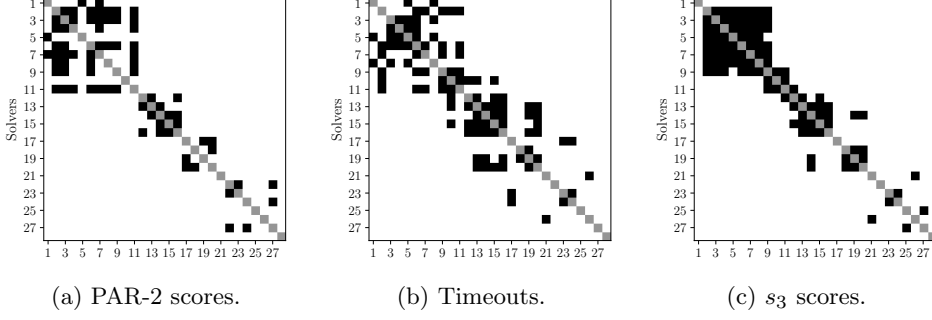


Fig. 3: Solver pairs with significant differences regarding the respective scorings are colored white. Non-significant differences are colored black. Solvers are ordered identically to Tab. 5.

5 Evaluation

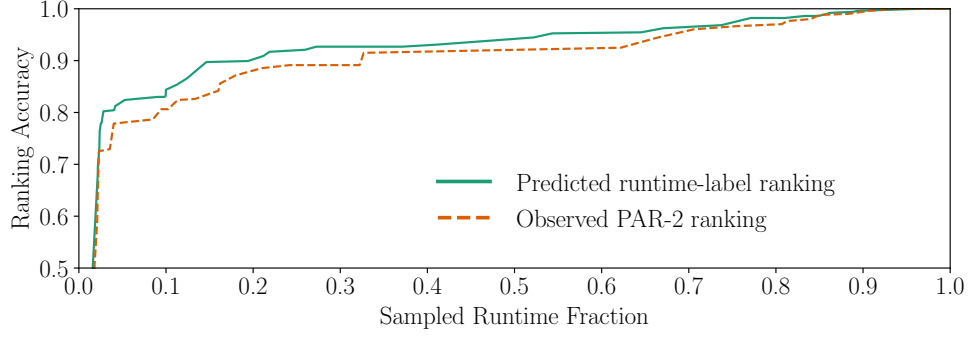
In this section, we evaluate our active learning framework. First, we analyze and tune the different subroutines of our framework on the tuning dataset. Next, we evaluate the best configurations on the full dataset. Finally, we analyze the importance of different instance families for our framework.

5.1 Hyperparameter Analysis

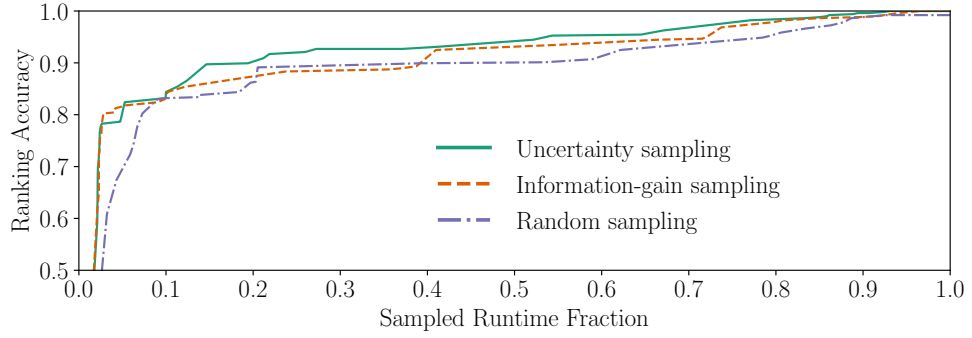
Our experiments follow the evaluation framework introduced in Section 4.1. Fig. 4 shows the performance of the approaches from Section 4.3 on O_{rt} - O_{acc} plots for the hyperparameter tuning dataset. Evaluating a given configuration with the algorithm 2 returns a point (O_{rt}, O_{acc}) . We do not show intermediate results of the active learning procedure, but only the final results after stopping. The plotted lines represent the best performing configurations per ranking approach (Fig. 4a), selection approach (Fig. 4b), and stopping criterion (Fig. 4c). In particular, we show the Pareto front, i.e., from all configurations that share a particular value of the plotted hyperparameter, we take the maximum ranking accuracy over all remaining hyperparameters *not* shown in the corresponding plot.

Regarding ranking approaches (Fig. 4a), using the predicted s_3 -induced runtime-label ranking consistently outperforms the partially observed PAR-2 ranking for every possible value of the tradeoff parameter δ . This result is expected, since selection decisions are not random. For example, we might sample more instances of a family if it benefits the discrimination of solvers. While the partially observed PAR-2 score is biased, the prediction model can account for this.

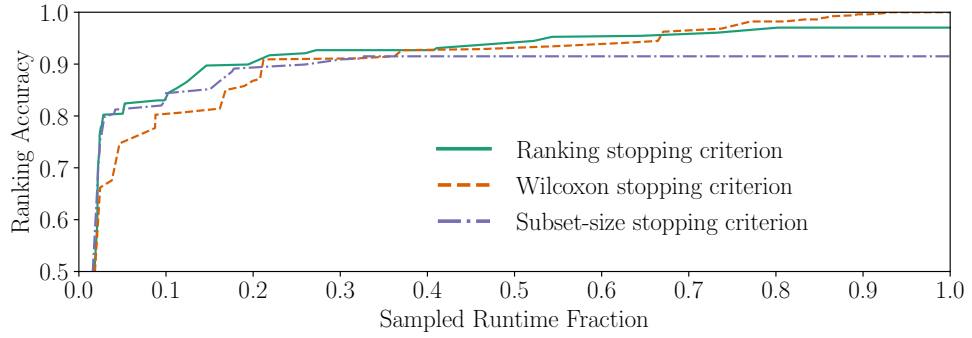
Regarding the selection approaches (Fig. 4b), uncertainty sampling performs best in most cases. However, information-gain sampling is advantageous when runtime is strongly favored (small δ ; runtime fraction less than 5%). This result is consistent with our expectations: Information-gain sampling selects instances that maximize the expected reduction in entropy. This means that we sample instances that reveal similarities between solvers rather than differences, which helps to build a confident model



(a) Ranking approaches



(b) Selection approaches



(c) Stopping criteria

Fig. 4: O_{rt} - O_{acc} -diagrams comparing different hyperparameter instantiations of our active-learning framework on the hyperparameter-tuning dataset. The x-axis shows the ratio of total solver runtime on the sampled instances relative to all instances. The y-axis shows the ranking accuracy (cf. Section 4.1). Each line entails the front of Pareto-optimal configurations for the respective hyperparameter instantiation.

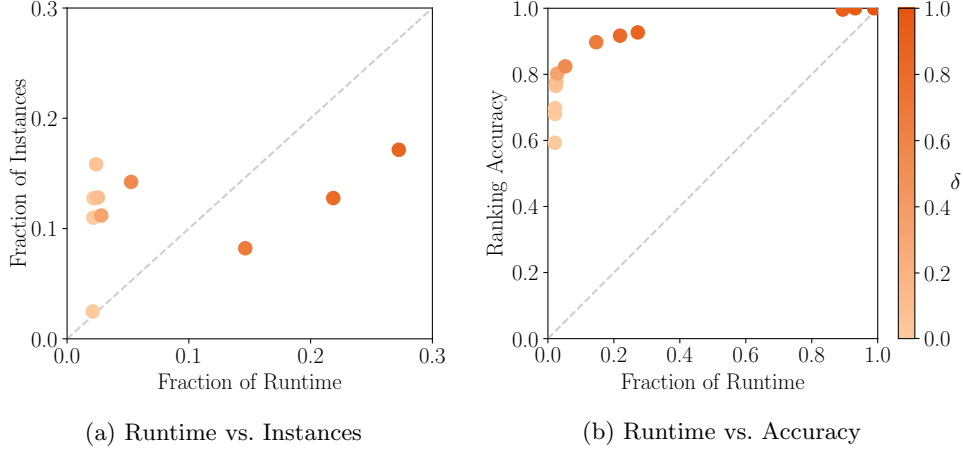


Fig. 5: Scatterplot comparing different instantiations of the tradeoff parameter δ for our active learning framework on the hyperparameter tuning dataset. The x-axis shows the fraction of runtime O_{rt} of the sample, while the y-axes show the fraction of sampled instances and the ranking accuracy, respectively. The color indicates the weighting between different optimization objectives $\delta \in [0, 1]$. The larger δ , the more we favor accuracy over runtime.

quickly. However, the method cannot select instances that are useful for distinguishing solvers later. Random sampling performs reasonably well, but is outperformed by uncertainty sampling in all cases, demonstrating the benefit of actively selecting instances based on a prediction model.

Regarding the stopping criteria (Fig. 4c), the ranking stopping criterion performs most consistently well. When accuracy is strongly favored (very high δ), the Wilcoxon stopping criterion performs better. The subset-size stopping criterion performs reasonably well, but does not improve beyond a certain accuracy because it samples a fixed subset of instances.

Fig. 5a shows an interesting consequence of weighting our optimization goals: On the one hand, if we want to get a *rough* estimate of a solver’s performance quickly (low δ), the approaches favor selecting many *easy* instances. In particular, the fraction of sampled instances is larger than the fraction of runtime. By having many observations, it is easier to build a model. On the other hand, if we want to get a *good* estimate of a solver’s performance in a moderate amount of time (high δ), approaches favor selecting few, *difficult* instances. In particular, the fraction of instances is smaller than the fraction of runtime. Furthermore, Fig. 5b shows which values of δ make the most sense. The range $\delta \in [0.2, 0.8]$ corresponds to the points with a runtime fraction between 0.03 and 0.22. We consider this region to be the most promising, analogous to the *elbow* method in cluster analysis [18].

Table 6: Performance comparison (on the full dataset) of the best performing AL approach (*AL*), random sampling of the same runtime fraction with 1000 repetitions (*Random*), and static selection of the instances most frequently sampled by active learning approaches (*Most Frequent*).

For $\delta \in [0.2, 0.7]$	AL	Random	Most Frequent
Sampled Runtime Fraction (%)	5.41	5.43	5.44
Sampled Instance Fraction (%)	26.53	5.43	27.75
Ranking Accuracy (%)	90.48	88.54	81.08
For $\delta \in (0.7, 0.8]$	AL	Random	Most Frequent
Sampled Runtime Fraction (%)	10.35	10.37	10.37
Sampled Instance Fraction (%)	5.24	10.37	36.96
Ranking Accuracy (%)	92.33	91.61	84.52

5.2 Full-Dataset Evaluation

After selecting the most promising hyperparameters, we run our active learning experiments on the complete Anniversary Track dataset (5355 instances). The above range $\delta \in [0.2, 0.8]$ results in only two different configurations. The best performing approach for $\delta \in [0.2, 0.7]$ uses the predicted runtime label ranking, information-gain sampling, and ranking stopping criterion. It can predict the PAR-2 ranking of a new solver with 90.48 % accuracy (O_{acc}) in only 5.41 % of the total evaluation time (O_{rt}). The best performing approach for $\delta \in (0.7, 0.8)$ uses the predicted runtime label ranking, uncertainty sampling, and ranking stopping criterion. It can predict the PAR-2 ranking of a new solver with 92.33 % accuracy (O_{acc}) in only 10.35 % of the total evaluation time (O_{rt}).

Table 6 shows how the two active learning approaches (column *AL*) compare to two static baselines: *Random* samples instances until it reaches roughly the same fraction of runtime as the AL benchmark sets. We repeat the sampling 1000 times and report the average results. *Most Frequent* uses a static benchmark set consisting of the instances most frequently sampled by our active learning approach. In particular, we consider the average sampling frequency across all solvers and Pareto-optimal active learning approaches.

Both of our AL approaches outperform random sampling. However, the performance differences are not significant with respect to a Wilcoxon signed-rank test with $\alpha = 0.05$, and also depend on the fraction of the runtime sampled (see Fig. 4b). However, a clear advantage of our approach is that it indicates when to stop adding new instances, depending on the tradeoff parameter δ . While the active learning results on the full dataset are not as strong as on the smaller tuning dataset, they still show the benefit of making the benchmark selection dependent on the solvers to be distinguished.

A static benchmark utilizing the most frequently AL-sampled instances exhibits suboptimal performance when compared to active learning and random sampling. This outcome is somewhat anticipated, given that the static benchmark fails to reflect an appropriate balance of instance families. For instance, families whose instances are uniformly randomly selected by AL, e.g., for different solvers, appear less frequently

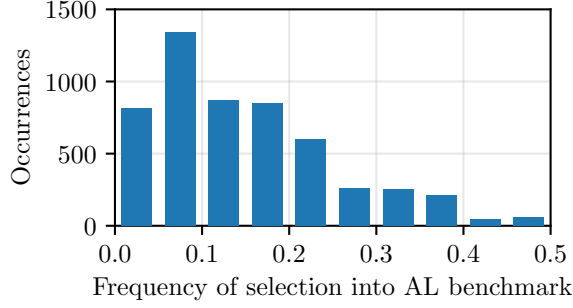


Fig. 6: The frequency with which individual problem instances are selected within all active-learning benchmarks provides insight into the relative importance of these instances. Instances that are chosen more frequently are intuitively more important for distinguishing solver runtimes.

in this benchmark than families where some instances are sampled more often than others.

5.3 Instance Frequency and Instance-Family Importance

Figure 6 illustrates the distribution of the frequency with which individual instances are selected across all active learning benchmarks. It can be observed that a small proportion of instances is selected with great frequency, while the majority of instances are selected on only a few occasions. This indicates that there is no fixed subset of instances that can perfectly distinguish all solvers, which highlights the value of employing an active learning-based strategy.

The selection decisions of our approach also reveal the importance of different instance families to our framework. Fig. 7 illustrates the occurrence of instance families within the dataset and the benchmarks created by active learning. We utilize the most effective configurations for all values of $\delta \in [0, 1]$ and examine the selection decisions made by the active learning approach on the SAT Competition 2022 Anniversary Track dataset [2]. While the majority of families exhibit the same fraction in the dataset and the sampled benchmarks, a few outliers require further discussion. Instances of the families *fpga*, *quasigroup-completion*, and *planning* are particularly useful for distinguishing solvers within our framework. These instances are selected in greater proportion than in the full dataset. In contrast, instances of the largest family, i.e., *hardware-verification*, appear with approximately the same fraction in the dataset and the sampled benchmarks. Finally, instances of the family *cryptography* are less important in distinguishing solvers than their substantial representation in the dataset would suggest. One possible explanation for this discrepancy is that these instances are highly similar, such that a small fraction of them is sufficient to estimate a solver’s performance on all of them.

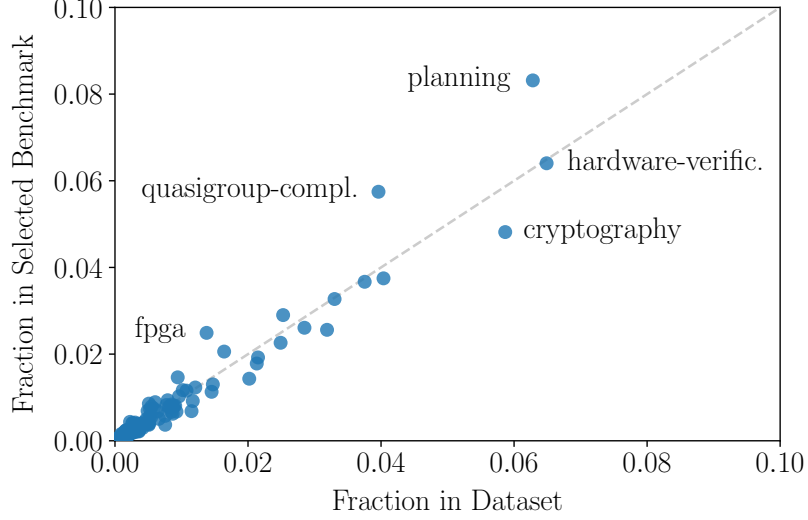


Fig. 7: The scatter plot illustrates the relative importance of different instance families for ranking. The x-axis depicts the frequency of instance families in the dataset, while the y-axis represents the average frequency of instance families in the samples selected by active learning. The dashed line represents families that occur with the same frequency in the dataset and samples.

6 Conclusions and Future Work

In this work, we have addressed the *New-Solver Problem*: Given a new solver, we aim to determine its ranking relative to competitors. Our approach provides accurate ranking predictions while requiring significantly less runtime than a complete evaluation on a given benchmark set. On data from the SAT Competition 2022 Anniversary Track, we can determine a new solver’s PAR-2 ranking with about 92 % accuracy while only requiring 10 % of the full-evaluation time. We have evaluated several ranking algorithms, instance-selection approaches, and stopping criteria within our sequential active learning framework. We have also looked at which instance families are the most prevalent in selection decisions.

Future work could compare more sub-routines for ranking, instance selection, and stopping. Our evaluation framework can be used for other computation-intensive problems, like other \mathcal{NP} -complete problems than SAT. Such problems share most of the relevant properties of SAT, like established instance features, a complete benchmark is expensive, and traditional benchmark selection requires expert knowledge.

From a technical perspective, we could formulate the problem of runtime discretization as an optimization problem, rather than addressing it empirically. Further, a major shortcoming of our current approach is the lack of parallelization, selecting instances one at a time. Benchmarking on a computing cluster with n cores benefits from having batches of n instances. However, bigger batch sizes n impede *active*

learning. Also, it is unclear how to synchronize instance selection and updates of the prediction model without wasting too much runtime.

In order to perform benchmarking on a computer cluster with multiple cores, it is necessary to have batches of multiple instances. This application demonstrates one of the principal limitations of our approach: it does not permit the concurrent execution of multiple instances, but rather necessitates the benchmarking of a single instance at a time. In future work, we intend to investigate models that learn from partial labels, which can be extracted from larger batch sizes in a parallel approach that synchronizes instance selection and model updates.

Declarations

Funding. This work was supported by the Ministry of Science, Research and the Arts Baden-Württemberg, project *Algorithm Engineering for the Scalability Challenge (AESC)*.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Availability of data and materials. All experimental data are available online at <https://github.com/mathefuchs/al-for-sat-solver-benchmarking-data>.

Code availability. The code is available online at <https://github.com/mathefuchs/al-for-sat-solver-benchmarking>.

References

- [1] Balint A, Belov A, Jarvisalo M, et al (2015) Overview and analysis of the SAT Challenge 2012 solver competition. *Artif Intell* 223:120–155. <https://doi.org/10.1016/j.artint.2015.01.002>
- [2] Balyo T, Heule M, Iser M, et al (eds) (2022) *Proceedings of SAT Competition 2022: Solver and Benchmark Descriptions*, Department of Computer Science, University of Helsinki, URL <http://hdl.handle.net/10138/347211>
- [3] Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- [4] Collautti M, Malitsky Y, Mehta D, et al (2013) SNNAP: solver-based nearest neighbor for algorithm portfolios. In: *Proc. ECML PKDD*, pp 435–450, https://doi.org/10.1007/978-3-642-40994-3_28
- [5] Dang N, Akgün Ö, Espasa J, et al (2022) A framework for generating informative benchmark instances. In: *Proc. CP*, pp 18:1–18:18, <https://doi.org/10.4230/LIPIcs.CP.2022.18>
- [6] De Winter JCF, Gosling SD, Potter J (2016) Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial

- using simulations and empirical data. *Psychol Methods* 21(3):273–290. <https://doi.org/10.1037/met0000079>
- [7] Dehghani M, Tay Y, Gritsenko AA, et al (2021) The benchmark lottery. arXiv:210707002 [csLG] URL <https://arxiv.org/abs/2107.07002>
 - [8] Froleys N, Heule M, Iser M, et al (2021) SAT Competition 2020. *Artif Intell* 301. <https://doi.org/10.1016/j.artint.2021.103572>
 - [9] Fuchs T, Bach J, Iser M (2023) Active learning for SAT solver benchmarking. In: *Proc. TACAS*, pp 407–425, https://doi.org/10.1007/978-3-031-30823-9_21
 - [10] Garzón I, Mesejo P, Giráldez-Cru J (2022) On the performance of deep generative models of realistic SAT instances. In: *Proc. SAT*, pp 3:1–3:19, <https://doi.org/10.4230/LIPIcs.SAT.2022.3>
 - [11] Gelder AV (2011) Careful ranking of multiple solvers with timeouts and ties. In: *Proc. SAT*, pp 317–328, https://doi.org/10.1007/978-3-642-21581-0_25
 - [12] Golbandi N, Koren Y, Lempel R (2011) Adaptive bootstrapping of recommender systems using decision trees. In: *Proc. WSDM*, pp 595–604, <https://doi.org/10.1145/1935826.1935910>
 - [13] Harpale A, Yang Y (2008) Personalized active learning for collaborative filtering. In: *Proc. SIGIR*, pp 91–98, <https://doi.org/10.1145/1390334.1390352>
 - [14] Hoos HH, Kaufmann B, Schaub T, et al (2013) Robust benchmark set selection for boolean constraint solvers. In: *Proc. LION*, pp 138–152, https://doi.org/10.1007/978-3-642-44973-4_16
 - [15] Hoos HH, Hutter F, Leyton-Brown K (2021) Automated configuration and selection of SAT solvers. In: *Handbook of Satisfiability*, 2nd edn. IOS Press, chap 12, p 481–507, <https://doi.org/10.3233/FAIA200995>
 - [16] Hutter F, Hoos HH, Leyton-Brown K (2011) Sequential model-based optimization for general algorithm configuration. In: *Proc. LION*, pp 507–523, https://doi.org/10.1007/978-3-642-25566-3_40
 - [17] Iser M, Sinz C (2018) A problem meta-data library for research in SAT. In: *Proc. PoS*, pp 144–152, <https://doi.org/10.29007/gdbb>
 - [18] Kodinariya TM, Makwana PR (2013) Review on determining number of cluster in k-means clustering. *Int J Adv Res Comput Sci Manage Stud* 1(6):90–95. URL <http://www.ijarcsms.com/docs/paper/volume1/issue6/V1I6-0015.pdf>
 - [19] Koren Y, Bell RM, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37. <https://doi.org/10.1109/MC.2009.263>

- [20] Manthey N, Möhle S (2016) Better evaluations by analyzing benchmark structure. In: Proc. PoS, URL http://www.pragmaticsofsat.org/2016/reg/POS-16-paper_4.pdf
- [21] Matricon T, Anastacio M, Fijalkow N, et al (2021) Statistical comparison of algorithm performance through instance selection. In: Proc. CP, pp 43:1–43:21, <https://doi.org/10.4230/LIPIcs.CP.2021.43>
- [22] Mısırl M (2017) Data sampling through collaborative filtering for algorithm selection. In: Proc. IEEE CEC, pp 2494–2501, <https://doi.org/10.1109/CEC.2017.7969608>
- [23] Mısırl M (2021) Benchmark set reduction for cheap empirical algorithmic studies. In: Proc. IEEE CEC, pp 871–877, <https://doi.org/10.1109/CEC45853.2021.9505012>
- [24] Mısırl M, Sebag M (2017) ALORS: An algorithm recommender system. Artif Intell 244:291–314. <https://doi.org/10.1016/j.artint.2016.12.001>
- [25] Ngoko Y, Cérlin C, Trystram D (2019) Solving SAT in a distributed cloud: A portfolio approach. Int J Appl Math Comput Sci 29(2):261–274. <https://doi.org/10.2478/amcs-2019-0019>
- [26] Nießl C, Herrmann M, Wiedemann C, et al (2022) Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. WIREs Data Min Knowl Discov 12(2). <https://doi.org/10.1002/widm.1441>
- [27] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12(85):2825–2830. URL <http://jmlr.org/papers/v12/pedregosa11a.html>
- [28] Rubens N, Elahi M, Sugiyama M, et al (2015) Active learning in recommender systems. In: Recommender Systems Handbook, 2nd edn. Springer, chap 24, p 809–846, https://doi.org/10.1007/978-1-4899-7637-6_24
- [29] Settles B (2009) Active learning literature survey. Tech. rep., University of Wisconsin-Madison, Department of Computer Sciences, URL <http://digital.library.wisc.edu/1793/60660>
- [30] Sinha S, Ebrahimi S, Darrell T (2019) Variational adversarial active learning. In: Proc. ICCV, pp 5971–5980, <https://doi.org/10.1109/ICCV.2019.00607>
- [31] Stützle T, López-Ibáñez M, Pérez-Cáceres L (2022) Automated algorithm configuration and design. In: Proc. GECCO, pp 997–1019, <https://doi.org/10.1145/3520304.3533663>

- [32] Tharwat A (2016) Linear vs. quadratic discriminant analysis classifier: a tutorial. *Int J Appl Pattern Recognit* 3(2):145–180. <https://doi.org/10.1504/IJAPR.2016.079050>
- [33] Tran T, Do T, Reid ID, et al (2019) Bayesian generative active deep learning. In: *Proc. ICML*, pp 6295–6304, URL <http://proceedings.mlr.press/v97/tran19a.html>
- [34] Volpato R, Song G (2019) Active learning to optimise time-expensive algorithm selection. *arXiv:1909.03261* [csLG] URL <https://arxiv.org/abs/1909.03261>
- [35] Wolpert DH (1992) Stacked generalization. *Neural Networks* 5(2):241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- [36] Xu L, Hutter F, Hoos HH, et al (2008) SATzilla: Portfolio-based algorithm selection for SAT. *J Artif Intell Res* 32:565–606. <https://doi.org/10.1613/jair.2490>
- [37] Xu L, Hutter F, Hoos HH, et al (2012) Features for SAT. Tech. rep., University of British Columbia, URL https://www.cs.ubc.ca/labs/beta/Projects/SATzilla/Report_SAT_features.pdf