Performance Prediction and Automated Tuning of Randomized and Parametric Algorithms

Frank Hutter[1], Youssef Hamadi[2], Holger H. Hoos[1], and Kevin Leyton-Brown[1]

[1] University of British Columbia, 2366 Main Mall, Vancouver BC, V6T1Z4, Canada
{hutter,kevinlb,hoos}@cs.ubc.ca
[2] Microsoft Research, 7 JJ Thomson Ave, Cambridge, UK
youssefh@microsoft.com

Abstract. Machine learning can be utilized to build models that predict the run-time of search algorithms for hard combinatorial problems. Such *empirical hardness models* have previously been studied for complete, deterministic search algorithms. In this work, we demonstrate that such models can also make surprisingly accurate predictions of the run-time distributions of incomplete and randomized search methods, such as stochastic local search algorithms. We also show for the first time how information about an algorithm's parameter settings can be incorporated into a model, and how such models can be used to automatically adjust the algorithm's parameters on a per-instance basis in order to optimize its performance. Empirical results for Novelty$^+$ and SAPS on structured and unstructured SAT instances show very good predictive performance and significant speedups of our automatically determined parameter settings when compared to the default and best fixed distribution-specific parameter settings.

1 Introduction

The last decade has seen a dramatic rise in our ability to solve combinatorial optimization problems in many practical applications. High-performance heuristic algorithms increasingly exploit problem instance structure. Thus, knowledge about the relationship between this structure and algorithm behavior forms an important basis for the development and successful application of such algorithms. This has inspired a large amount of research on methods for extracting and acting upon such information. These range from search space analysis to automated algorithm selection and tuning methods.

An increasing number of studies explore the use of machine learning techniques in this context [15, 18, 7, 5, 8]. One recent approach uses linear basis function regression to obtain models of the time an algorithm will require to solve a given problem instance [19, 22]. These so-called *empirical hardness models* can be used to obtain insights into the factors responsible for an algorithm's performance, or to induce distributions of problem instances that are challenging for a given algorithm. They can also be leveraged to select among several different algorithms for solving a given problem instance.

In this paper, we extend on this work in three significant ways. First, past work on empirical hardness models has focused exclusively on complete, deterministic algorithms. Our first goal is to show that the same approach can be used to predict sufficient statistics of the run-time distributions (RTDs) of incomplete, randomized algorithms,

and in particular of stochastic local search (SLS) algorithms for SAT. This is important because SLS algorithms are among the best existing techniques for solving a wide range of hard combinatorial problems, including hard subclasses of SAT [14].

The behavior of many randomized heuristic algorithms is controlled by parameters with continuous or large discrete domains. This holds in particular for most state-of-the-art SLS algorithms. For example, the performance of WalkSAT algorithms such as Novelty [20] or Novelty$^+$ [12] depends critically on the setting of a noise parameter whose optimal value is known to depend on the given SAT instance [13]. Understanding the relationship between parameter settings and the run-time behavior of an algorithm is of substantial interest for both scientific and pragmatic reasons, as it can expose weaknesses of a given search algorithm and help to avoid the detrimental impact of poor parameter settings. Thus, our second goal is to extend empirical hardness models to include algorithm parameters in addition to features of the given problem instance.

Finally, hardness models could also be used to automatically determine good parameter settings. Thus, an algorithm's performance could be optimized for each problem instance without any human intervention or significant overhead. Our final goal is to explore the potential of such an approach for automatic per-instance parameter tuning.

In what follows, we show that we have achieved all three of our goals by reporting the results of experiments with SLS algorithms for SAT.[3] Specifically, we considered two high-performance SLS algorithms for SAT, Novelty$^+$ [12] and SAPS [17], and several widely-studied structured and unstructured instance distributions. In section 2, we show how to build models that predict the sufficient statistics of RTDs for randomized algorithms. Empirical results demonstrate that we can predict the median run-time for our test distributions with surprising accuracy (we achieve correlation coefficients between predicted and actual run-time of up to 0.995), and that based on this statistic we can also predict the complete exponential RTDs Novelty$^+$ and SAPS exhibit. Section 3 describes how empirical hardness models can be extended to incorporate algorithm parameters; empirical results still demonstrate excellent performance for this harder task (correlation coefficients reach up to 0.98). Section 4 shows that these models can be leveraged to perform automatic per-instance parameter tuning that results in significant reductions of the algorithm's run-time compared to using default settings (speedups of up to two orders of magnitude) or even the best fixed parameter values for the given instance distribution (speedups of up to an order of magnitude). Section 5 describes how Bayesian techniques can be leveraged when predicting run-time for test distributions that differ from the one used for training of the empirical hardness model. Finally, Section 6 concludes the paper and points out future work.

2 Run-time Prediction: Randomized Algorithms

Previous work [19, 22] has shown that it is possible to predict the run-time of deterministic tree-search algorithms for combinatorial problems using supervised machine

[3] We should note that our approach is not limited to SLS algorithms (it applies as well to e.g., randomized tree-search methods). Furthermore our techniques are not especially tuned to the SAT domain, though the features we use were created with some domain knowledge. In experimental work it is obviously necessary to choose *some* specific domain. We have chosen to study the SAT problem because it is the prototypical and best-studied $\mathcal{NP}$-complete problem and there exists a great variety of SAT benchmark instances and solvers.

learning techniques. In this section, we demonstrate that similar techniques are able to predict the run-time of algorithms which are both randomized and incomplete. We support our arguments by presenting the results of experiments involving two competitive local search algorithms for SAT.

2.1 Prediction of sufficient statistics for run-time distributions

It has been shown in the literature that high-performance randomized local search algorithms tend to exhibit exponential run-time distributions [14], meaning that the run-times of two runs that differ only in their random seeds can easily vary by more than one order of magnitude. Even more extreme variability in run-time has been observed for randomized complete search algorithms [10]. Due to this inherent randomness of the algorithm (and since we do not incorporate information on a particular run), we have to predict a run-time distribution, that is, a probability distribution over the amount of time an algorithm will take to solve the problem. Run-time distributions for many randomized algorithms closely resemble standard parametric distributions, such as the exponential or the Weibull distribution (see, e.g., [14]). These parametric distributions are completely specified by a small number of sufficient statistics; for example, an exponential distribution is completely specified by its median (or any other quantile) or its mean. It follows that by predicting these sufficient statistics, a prediction for the entire run-time distribution for an unseen instance can be obtained.

Note that for randomized algorithms, the error in a model's predictions can be thought of as consisting of two components, one of which describes the extent to which the model fails to describe the data, and the second of which expresses the inherent noise in the employed summary statistics due to randomness of the algorithm. This latter component reduces as we increase the number of runs from which we extract the statistic. We demonstrate this later in this section (Figures 1(a) and 1(b)): while empirical hardness models of SLS algorithms are able to predict the run-times of single runs reasonably well, their predictions of median run-times over a larger set of runs are much more accurate.

Our approach for run-time prediction of randomized incomplete algorithms largely follows the linear regression approach of [19]. We have previously explored other techniques, namely support vector machine regression, multivariate adaptive regression splines (MARS), and lasso regression, none of which improved our results significantly.[4] While we can handle both complete and incomplete algorithms, we restrict our experiments to incomplete local search algorithms. We note, however, that an extension of our work to randomized tree search algorithms would be straightforward.

In order to predict the run-time of an algorithm $\mathcal{A}$ on a distribution $\mathcal{D}$ of instances, we draw an i.i.d. sample of N instances from $\mathcal{D}$. For each instance s_n in this training set, $\mathcal{A}$ is run some constant number of times and an empirical fit r_n of the sufficient statistics of interest is recorded. Note that r_n is a $1 \times M$ vector if there are M sufficient statistics of interest. We also compute a set of 46 instance features $\boldsymbol{x}_n = [x_{n,1}, \ldots, x_{n,k}]$ for each instance. This set is a subset of the features used in [22], including basic statistics,

[4] Preliminary experiments (see Section 5) suggest that Gaussian processes can increase performance, especially when the amount of available data is small, but their cubic scaling behavior in the number of data points complicates their use in practice.

graph-based features, local search probes, and DPLL-based measures.[5] We restricted the subset of features because some features timed out for large instances – the computation of all our 46 features took just over 2 seconds for each instance.

Given this data for all the training instances, a function $f(\boldsymbol{x})$ is fitted that, given the features $\boldsymbol{x}_n$ of an instance, approximates $\mathcal{A}$'s *median* run-time r_n. In linear basis function regression, we construct this function by first generating a set of more expressive basis functions $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\boldsymbol{x}_n) = [\phi_1(\boldsymbol{x}_n), \ldots, \phi_D(\boldsymbol{x}_n)]$ which can include arbitrarily complex functions of *all* features $\boldsymbol{x}_n$ of an instance s_n, or simply the raw features themselves. These basis functions typically contain a number of elements which are either unpredictive or highly correlated, so this set should be reduced by applying some form of feature selection. We then use ridge regression to fit the $D \times M$ matrix of free parameters $\boldsymbol{w}$ of a linear function $f_{\boldsymbol{w}}(\boldsymbol{x}_n) = \boldsymbol{\phi}(\boldsymbol{x}_n)^{\boldsymbol{T}}\boldsymbol{w}$, that is, we compute $\boldsymbol{w} = (\delta I + \boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{r}$, where δ is a small regularization constant (set to 10^{-2} in our experiments), $\boldsymbol{\Phi}$ is the $N \times D$ design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_N^T]^T$, and $\boldsymbol{r} = [{r_1}^T, \ldots, {r_N}^T]^T$. Given a new, unseen instance s_{N+1}, a prediction of the M sufficient statistics can be obtained by computing the instance features $\boldsymbol{x}_{N+1}$ and evaluating $f_{\boldsymbol{w}}(\boldsymbol{x}_{N+1}) = \boldsymbol{\phi}(\boldsymbol{x}_{N+1})^{\boldsymbol{T}}\boldsymbol{w}$. One advantage of the simplicity of ridge regression is a low computational complexity of $\Theta(\max\{D^3, D^2N, DNM\})$ for training and of $\Theta(DM)$ for prediction for an unseen test instance.

2.2 Experimental setup and empirical results for predicting median run-time

We performed experiments for the prediction of median run-time using two SLS algorithms, SAPS and Novelty$^+$. In this section we fix SAPS parameters to their defaults $\langle \alpha, \rho, P_{smooth} \rangle = \langle 1.3, 0.8, 0.05 \rangle$. For Novelty$^+$, we use its default parameter setting $\langle noise, wp \rangle = \langle 50, 0.01 \rangle$ for unstructured instances; for structured instances where Novelty$^+$ is known to perform better with lower noise settings, we chose $\langle noise, wp \rangle = \langle 10, 0.01 \rangle$ (with noise=50% the majority of runs did not finish within an hour of CPU time). We consider models that incorporate multiple parameter settings in the next section. Since, based on previous studies of these algorithms, we expect approximately exponential run-time distributions, the empirical fit of the sufficient statistics r_n for each training instance s_n simply consists of recording the median of a fixed number of runs.

In our experiments, we used six widely-studied SAT benchmark distributions, three of which consist of unstructured instances and the other three of structured instances. The first two distributions we studied each consisted of 20,000 uniform-random 3-SAT instances with 400 variables; the first (**CV-var**) varied the clauses-to-variables ratio between 3.26 and 5.26, while the second (**CV-fixed**) fixed $c/v = 4.26$. These distributions were previously studied in [22], facilitating a comparison of our results with past work. Our third unstructured distribution (**SAT04**) consisted of 3,000 random unstructured instances generated with the two generators used for the 2004 SAT solver competition (with identical parameters) and was employed to evaluate our automated parameter tuning procedure on a competition benchmark.

Our first two structured distributions are different variants of quasigroup completion problems. The first one (**QCP**) consisted of 30,000 hard quasigroup completion

[5] Information on precisely which features we used, as well as the rest of our experimental data and Matlab code, is available online at `http://www.cs.ubc.ca/labs/beta/Projects/RandomHardness/`.

Unstructured instances					
Dataset	N	Algorithm	Runs	Corrcoeff	RMSE
CV-var	10011	SAPS	1	0.892/0.906	0.38/0.36
CV-var	10011	SAPS	10	0.949/0.964	0.26/0.21
CV-var	10011	SAPS	100	0.959/0.972	0.23/0.19
CV-var	10011	SAPS	1000	0.958/0.975	0.23/0.18
CV-var	10011	Novelty$^+$	10	0.947/0.949	0.24/0.23
CV-fixed	10129	SAPS	10	0.758/0.784	0.45/0.43
CV-fixed	10129	Novelty$^+$	10	0.558/0.596	0.61/0.59
SAT04	1420	SAPS	10	0.918/0.910	0.55/0.55
SAT04	1420	Novelty$^+$	10	0.918/0.915	0.59/0.61

Structured instances					
Dataset	N	Algorithm	Runs	Corrcoeff	RMSE
QWH	7498	SAPS	10	0.983/0.991	0.38/0.28
QWH	9601	Novelty$^+$	10	0.990/0.993	0.25/0.21
QCP	8117	SAPS	10	0.995/0.996	0.17/0.16
QCP	14927	Novelty$^+$	10	0.994/0.995	0.11/0.11
SW-GCP	2740	SAPS	10	0.886/0.881	0.45/0.45
SW-GCP	11181	Novelty$^+$	10	0.747/0.751	0.23/0.23

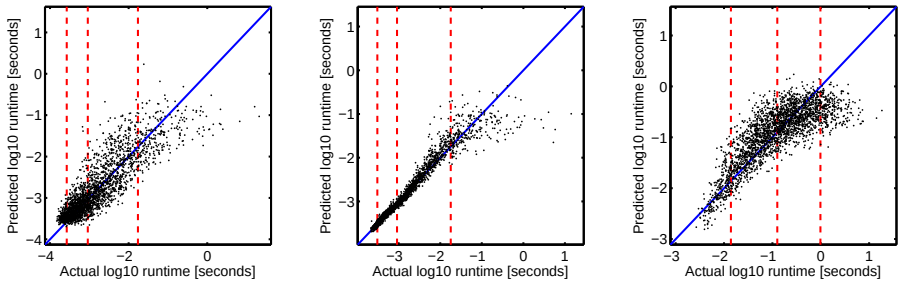
Table 1. Evaluation of learned models on test data. N is the number of instances for which the algorithm's median runtime is ≤ 900 CPU seconds (only those instances are used and split 50:25:25 into training, validation, and test sets). Columns for correlation coefficient and RMSE indicate values using only raw features as basis functions, and then using raw features and their pairwise products. SAPS was always run with its default parameter settings $\langle \alpha, \rho \rangle = \langle 1.3, 0.8 \rangle$. For Novelty$^+$, we used noise=50% for unstructured and noise=10% for structured instances.

problems, while the second one (**QWH**) contained 9,601 instances of the quasigroup completion problem for quasigroups with randomly punched holes) [11]. Both distributions were created with the generator `lsencode` by Carla Gomes. The ratio of unassigned cells varied from 25% to 75%. We chose quasigroup completion problems as a representative of structured problems because this domain allows the systematic study of a large instance set with a wide spread in hardness, and because the structure of the underlying Latin squares is similar to the one found in many real-world applications, such as scheduling, time-tabling, experimental design, and error correcting codes [11]. Our last structured distribution (**SW-GCP**) contained 20,000 instances of graph colouring based on small world graphs that were created with the generator `sw.lsp` by Toby Walsh [9].

As is standard in the study of SLS algorithms, all distributions were filtered to contain only satisfiable instances, leading to 10,011, 10,129, 1,420, 17,989, 9,601, and 11,181 instances for CV-var, CV-fixed, SAT04, QCP, QWH, and SW-GCP respectively.We then randomly split each instance set 50:25:25 into training, validation, and test sets; all experimental results are based on the test set and were stable with respect to reshuffling. We chose the 46 raw features and the constant 1 as our basis functions, and also included pairwise multiplicative combinations of all raw features. We then performed forward selection to select up to 40 features, stopping when the error on the validation set first began to grow. Experiments were run on a cluster of 50 dual 3.2GHz Intel Xeon PCs with 2MB cache and 2GB RAM, running SuSE Linux 9.1. All runs were cut off after 900 seconds, discarding the corresponding data points.

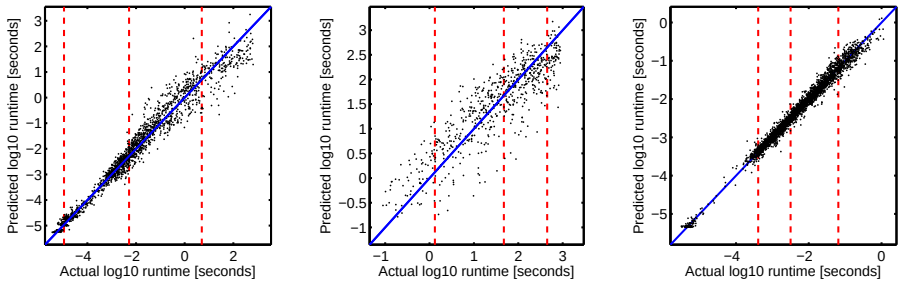
Overall, our experiments show that we can consistently predict median run-time with surprising accuracy. The results of experiments on our benchmark distributions are summarized in Table 1. Note that a correlation coefficient of 1 indicates perfect prediction while 0 indicates random noise; a root mean squared error (RMSE) of 0 means perfect prediction while 1 roughly means average misprediction by one order of magnitude. Also note that our predictions for Novelty$^+$ and SAPS are qualitatively similar.

Figure 1(a) shows a scatterplot of predicted vs. actual run-time for Novelty$^+$ on CV-var, where the model is trained and evaluated on a single run per instance. Most of the data points are located in the very left of this plot, which we visualize by plotting the 10%, 50% and 90% quantiles of the data (the three dotted vertical bars). While a



(a) 1 Novelty$^+$ run on CV-var. CC=0.880, RMSE=0.38 (b) 100 Novelty$^+$ runs on CV-var. CC=0.962, RMSE=0.20 (c) 100 SAPS runs on CV-fixed. CC=0.814, RMSE=0.40

Fig. 1. Correlation between observed and predicted run-times/medians of run-times of SAPS and Novelty$^+$ on unstructured instances. The basis functions were raw features and their pairwise products. The three vertical bars in these and all other scatter plots in this paper denote the 10%, 50%, and 90% quantiles of the data. For example, this means that 40% of the data points lie between the left and the middle vertical bar.



(a) 10 SAPS runs on QWH (b) 10 SAPS runs on SW-GCP (c) 10 Novelty$^+$ runs on QCP

Fig. 2. Correlation between observed and predicted run-times/medians of run-times of SAPS and Novelty$^+$ on SAT04 and QWH. The basis functions were raw features and their pairwise products. For RMSEs and correlations coefficients, see Table 1.

strong trend is evident in Figure 1(a), there is significant error in the predictions. Figure 1(b) shows the same algorithm on the same dataset, but now predicting the median of an empirical run-time distribution based on 100 runs of Novelty$^+$. The error for the leftmost 90% of the data points is substantially reduced, leading to an almost halved RMSE when compared to predictions for a single run. It is also noteworthy that these run-time predictions are more accurate than the predictions for the deterministic algorithms kcnfs, satz, and oksolver (compare against Figure 5(left) in [22]). While this is already true for predictions based on single runs it is much more pronounced when predicting median run-time. This same effect holds true for predicting median run-time of SAPS, and for different distributions. Figure 1(c) also shows much better predictions than we observed for deterministic tree search algorithms on CV-fix (compare this plot against Figure 7(left) in [22]).

We believe that two factors contribute to this effect. First, we see deterministic algorithms as roughly equivalent to randomized algorithms with a fixed seed. Obviously, the single run-time of such an algorithm on a particular instance is less informative about

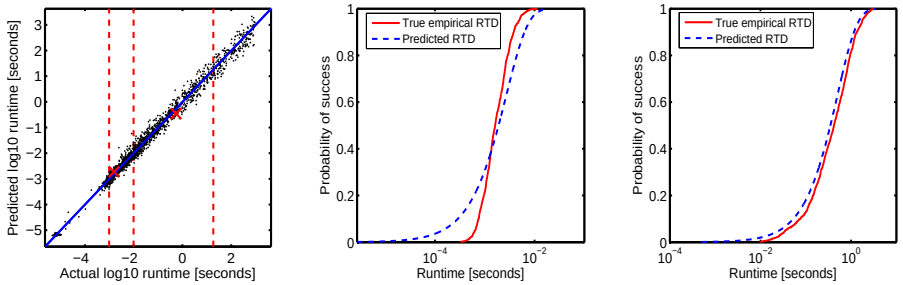
#	Basis function	Cost of omission	Corrcoeff	RMSE
	SAPS on CV-fix			
1.	saps_BestSolution_CoeffVariance × saps_BestStep_CoeffVariance	100	0.744/0.785	0.47/0.44
1.	saps_BestSolution_CoeffVariance × saps_AvgImproveToBest_Mean	100		
2.	saps_BestStep_CoeffVariance × saps_FirstLMRatio_Mean	45		
3.	gsat_BestSolution_CoeffVariance × lobjois_mean_depth_over_vars	37	0.758/0.785	0.46/0.44
4.	saps_AvgImproveToBest_CoeffVariance	15		
5.	saps_BestCV_Mean × gsat_BestStep_Mean	11		
	Novelty$^+$ on QCP			
1.	VG_mean × gsat_BestStep_Mean	100	0.966/0.994	0.29/0.11
1.	saps_AvgImproveToBest_CoeffVariance × gsat_BestSolution_Mean	100		
2.	vars_clauses_ratio × lobjois_mean_depth_over_vars	68		
3.	VG_mean × gsat_BestStep_Mean	12	0.991/0.994	0.13/0.11
4.	TRINARY_PLUS × lobjois_log_num_nodes_over_vars	7		

Table 2. Feature importance in small subset models for predicting median run-time of 10 runs. The cost of omission for a feature specifies how much worse validation set predictions are without it, normalized to 100 for the top feature. The RMSE and Corrcoeff columns compare predictive quality on the test set to that of full 40-feature models.

its underlying run-time distribution (were it randomized) than the sufficient statistics of multiple runs. We thus conjecture that it should be possible to achieve more accurate predictions of deterministic algorithms by randomizing them in a meaningful way. Second, one of the main reasons to introduce randomness in search is to achieve diversification. This allows the heuristic to recover from making a bad decision by exploring a new part of the search space, and hence reduces the variance of (e.g., median) run-times across very similar instances. Because deterministic solvers do not include such diversification mechanisms, they can exhibit strikingly different run-times on very similar instances. For example, consider modifying a SAT instance by randomly shuffling the names of its variables. One would expect a properly randomized algorithm to have virtually the same run-time distributions for the original and modified instances; however, a deterministic solver which breaks ties in its variable-ordering or local search heuristic according to a lexicographic ordering could exhibit very different runtimes on the two instances. Empirical hardness models must give similar predictions for instances with similar feature values. If for similar instances the similarity between run-time distributions of randomized algorithms is greater than the similarity between the run-times of deterministic solvers, one would expect empirical hardness models for randomized algorithms to be more accurate.

Figure 2 visualizes our predictive quality for structured data sets. Performance for both QWH and QCP, as shown in Figures 2(a) and 2(c), was outstanding with correlation coefficients between predicted and actual median run-time of up to 0.996. The last structured data set, SW-GCP, is the hardest distribution for prediction we have encountered thus far (unpublished work shows RMSEs of around 1.0 when predicting the run-time of deterministic algorithms on SW-GCP). As shown in Figure 2(b), the predictions for SAPS are surprisingly good; predictive quality for Novelty$^+$ (see Table 1) is also much higher than what we have seen for deterministic algorithms.

We now look at which features are most important to our models; this is not straightforward since the features are highly correlated. Following [19, 22], we build subset models of increasing size until the RMSE and correlation coefficient are comparable to the ones for the full model with 40 basis functions. Table 2 reports the results for SAPS on CV-fix and Novelty$^+$ on QCP and for each of these also gives the performance of



(a) Predictions of median run-time, instances $q_{0.25}$ and $q_{0.75}$ (b) Easy QCP instance ($q_{0.25}$) (c) Hard QCP instance ($q_{0.75}$)

Fig. 3. Predicted versus actual empirical RTDs for SAPS on two QCP instances. 10 runs were used for learning median run-time, 1000 runs to plot the empirical RTDs in (b) and (c).

the best model with a single basis function. Overall, we observe that the most important features for predicting run-time distributions of our SLS algorithms are the same ones that were observed to be important for predicting run-times of deterministic algorithms in [22]. Also similar to observations from [22], we found that very few features are needed to build run-time models of instances that are all satisfiable. While [22] studied only uniform-random data, we found in our experiments that this is true for both unstructured and structured instances and for both algorithms we studied. Small models for CV-var (both for SAPS and Novelty$^+$) almost exclusively use local search features (almost all of them based on short SAPS trajectories). The structured domain QCP employs a mix of local search probes (based on both SAPS and GSAT), constraint-graph-based features (e.g., VG_mean) and in the case of Novelty$^+$ also some DPLL-based features, such as the estimate of the search tree size (lobjois_mean_depth_over_vars).

In some cases (e.g., models of SAPS on CV-var), a single feature is already able to predict single run-times with virtually the same accuracy as the full model. However, the relative quality of these simple models degrades as we increase the number of runs over which the sufficient statistics are based. This makes intuitive sense when one realizes that sufficient statistics that are based on more runs are more stable and simply yield higher-quality data (less noisy observations); when the data is poor, complicated models do not pay off, but as data quality increases so does the performance margin between simple and complicated models.

To illustrate the fact that based on the median we can predict fairly accurately entire run-time distributions for the SLS algorithms studied here, we show the predicted and empirically measured RTDs for SAPS on two QCP instances in Figure 3. The two instances correspond to the 0.25 and 0.75 quantiles of the distribution of actual median hardness for SAPS on the entire QCP instance set; they correspond to the red crosses in Figure 3(a), which shows the tight correlation between actual and predicted run-times. Consistent with previous results by Hoos et al. (see, e.g., Chapters 4 and 6 of [14]), the RTD for the $q_{0.75}$ instance exhibits an exponential RTD, which our approach almost perfectly fits (see Figure 3(c)). The RTDs for easier instances are known to typically exhibit smaller variance; therefore, an approximation with an exponential distribution is less accurate (see Figure 3(b)). We plan to predict sufficient statistics for the more

general distributions needed to characterise such RTDs, such as Weibull and generalised exponential distributions, in the future.

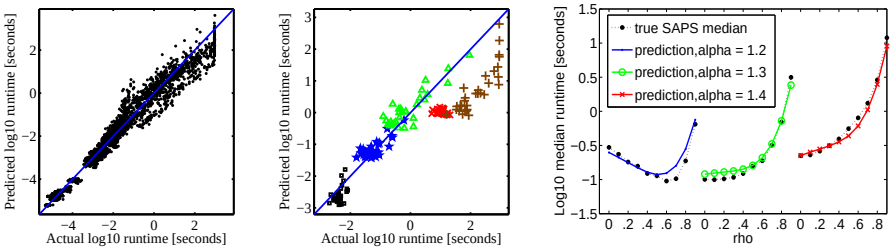
3 Run-time Prediction: Parametric Algorithms

It is well known that the behavior of most high-performance SLS algorithms is controlled by one or more parameters, and that these parameters often have a substantial effect on the algorithm's performance [14]. In the previous section, we showed that good empirical hardness models can be constructed when these parameters are held constant. In practice, however, we want to be able to change these parameter values and to understand what will happen when we do. In this section, we demonstrate that it is possible to incorporate parameters into empirical hardness models for randomized, incomplete algorithms. Our techniques should carry over directly to both deterministic and complete parametric algorithms (in the case of deterministic algorithms using single run-times instead of sufficient statistics of RTDs).

Our approach is to learn a function $g(\boldsymbol{x}, c)$ that takes as inputs both the features $\boldsymbol{x}_n$ of an instance s_n and the parameter configuration c of an algorithm $\mathcal{A}$, and that approximates sufficient statistics of $\mathcal{A}$'s RTD when run on instance s_n with parameter configuration c. In the training phase, for each training instance s_n we run $\mathcal{A}$ some constant number of times with a set of parameter configurations $\boldsymbol{c}_n = \{c_{n,1}, \ldots, c_{n,k_n}\}$, and collect fits of the sufficient statistics $\boldsymbol{r}_n = [r_{n,1}^T, \ldots, r_{n,k_n}^T]^T$ of the corresponding empirical run-time distributions. We also compute s_n's features $\boldsymbol{x}_n$. The key change from the approach in Section 2.1 is that now the parameters that were used to generate an $\langle$instance,run-time$\rangle$ pair are effectively treated as additional features of that training example. We define a new set of basis functions $\boldsymbol{\phi}(\boldsymbol{x}_n, c_{n,j}) = [\phi_1(\boldsymbol{x}_n, c_{n,j}), \ldots, \phi_D(\boldsymbol{x}_n, c_{n,j})]$ whose domain now consists of the cross product of features and parameter configurations. For each instance s_n and parameter configuration $c_{n,j}$, we will have a row in the design matrix $\boldsymbol{\Phi}$ that contains $\boldsymbol{\phi}(\boldsymbol{x}_n, c_{n,j})^T$ — that is, the design matrix now contains n_k rows for every training instance. The target vector $\boldsymbol{r} = [\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_N^T]^T$ just stacks all the sufficient statistics on top of each other. We learn the function $g_{\boldsymbol{w}}(\boldsymbol{x}, c) = \boldsymbol{\phi}(\boldsymbol{x}, c)^{\boldsymbol{T}} \boldsymbol{w}$ by applying ridge regression as in Section 2.1.

Our experiments in this section concentrate on predicting median run-time of SAPS since it has three interdependent, continuous parameters, as compared to Novelty$^+$ which has only two parameters, one of which (wp) is typically set to a default value that results in uniformly good performance. This difference notwithstanding, we observed qualitatively similar results with Novelty$^+$. Note that the approach outlined above allows one to use different parameter settings for each training instance. How to pick these parameter settings for training in the most informative way is an interesting experimental design which invites application of active learning techniques; we plan to tackle it in future work. In this study, for SAPS we used all combinations of $\alpha \in \{1.2, 1.3, 1.4\}$ and $\rho \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$, keeping $P_{smooth} = 0.05$ constant since its effect is highly correlated with that of ρ. For Novelty$^+$, we used $noise \in \{10, 20, 30, 40, 50, 60\}$, fixing $wp = 0.01$.

As basis functions, we used multiplicative combinations of the raw instance features $\boldsymbol{x}_n$ and a 2nd order expansion of all non-fixed (continuous) parameter settings. For K raw features ($K = 46$ in our experiments), this meant $3K$ basis functions for Novelty$^+$,



(a) SAPS on QWH, 30 settings (b) 5 symbol-coded instances (c) Run-time predictions for median-hard instance

Fig. 4. (a) Predictions for SAPS on QWH with 30 parameter settings. (b) Data points for 5 instances from SAPS on SAT04, different symbol for each instance. (c) Predicted run-time vs. median SAPS run-time over 1000 runs for 30 parameter settings on the median SAT04 instance (in terms of hardness for SAPS with default parameter settings). The learnt model uses a 4th order expansion of the parameters, multiplicatively combined with the instance features.

and $6K$ for SAPS, respectively. As before we applied forward selection to select up to 40 features, stopping when the error on the validation set first began to grow. For each data set reported here, we randomly picked 1000 instances to be split 50:50 for training and validation. We ran one run per instance and parameter configuration yielding 30,000 data points for SAPS and 6,000 for Novelty$^+$. (Training on the median of more runs would likely have improved the results.) For the test set, we used an additional 100 distinct instances and computed the median of 10 runs for each parameter setting.

In Figure 4(a), we show predicted vs. actual SAPS run-time for the QWH dataset, where the SAPS parameters are varied as described above. This may be compared to Figure 2(a), which shows the same algorithm on the same dataset for fixed parameter values. (Note, however, that Figure 2(a) was trained on more runs and using more powerful basis functions for the instance features.) We observe that our model still achieves excellent performance, yielding a correlation coefficient of .98 and an RMSE of .40, as compared to .98 and .38 respectively for the fixed-parameter setting (using raw features as basis functions); for Novelty$^+$, the numbers are .98 and .58, respectively.

Figure 4(b) shows predicted vs. actual SAPS median run-time for five instances from SAT04, namely the easiest and hardest instance, and the 25%, 50%, and 75% quantiles. Runs corresponding to the same instance are plotted using the same symbol. Observe that run-time variation due to the instance is often greater than variation due to parameter settings. However, harder instances tend to be more sensitive to variation in the algorithm's parameters than easier ones – this indicates the importance of parameter tuning, especially for hard instances. The average correlation coefficient for the 30 points per instance is .52; for the 6 points per instance in Novelty$^+$ it is .86, much higher.

Figure 4(c) shows SAPS run-time predictions for the median instance of our SAT04 test set at each of its 30 $\langle \alpha, \rho \rangle$ combinations; these are compared to the actual median SAPS run-time on this instance. We observe that the learned model very closely predicts the actual run-times, despite the fact that the relationship between run-time and the two parameters is complex. In the experiment upon which Figure 4(c) is based feature selection chose 40 features; thus, the model learned a 40-dimensional surface and the figure shows that its projection onto the 2-dimensional parameter space at the current instance features closely matches the actual run-time for this instance.

Set	Algo	Gross corr	RMSE	Corr per inst.	best fixed params	s_{bpi}	s_{wpi}	s_{def}	s_{fixed}
SAT04	Nov	0.90	0.78	0.86	50	0.62	275.42	0.89	0.89
QWH	Nov	0.98	0.58	0.69	10	0.81	457.09	**177.83**	0.91
Mixed	Nov	0.95	0.8	0.81	40	0.74	363.08	**13.18**	**10.72**
SAT04	SAPS	0.95	0.67	0.52	$\langle 1.3, 0\rangle$	0.56	10.72	**2.00**	**1.07**
QWH	SAPS	0.98	0.40	0.39	$\langle 1.2, .1\rangle$	0.65	6.03	**2.00**	0.93
Mixed	SAPS	0.91	0.60	0.65	$\langle 1.2, 0.2\rangle$	0.46	17.78	**1.91**	0.93

Table 3. Results for automated parameter tuning. For each instance set and algorithm, we give the correlation between actual and predicted run-time for all instances, RMSE, the average correlation for all the data points of an instance, and the best fixed parameter setting on the test set. We also give the average speedup over the best possible parameter setting per instance (s_{bpi}, always ≤ 1), over the worst possible setting per instance (s_{wpi}, always ≥ 1), the default (s_{def}), and the best fixed setting on the test set. For example, on Mixed, Novelty$^+$ is on average 10.72 times faster than its best data-set specific fixed parameter setting (s_{fixed}). All experiments use second order expansions of the parameters (combined with the instance features). Bold face indicates speedups of the automated parameter setting over the default and best fixed parameter settings.

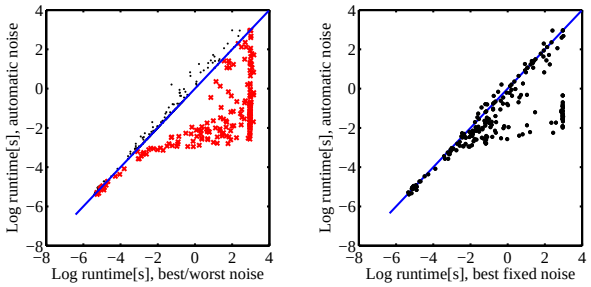
4 Automated Parameter Tuning

Our results, as suggested by Figures 4(a) and 4(b), indicate that our methods are able to predict per-instance and per-parameter run-times with high accuracy. We can thus hope that they would also be able to accurately predict which parameter settings result in the lowest run-time for a given instance. This would allow us to use a learned model to automatically tune the parameter values of an SLS algorithm on a per-instance basis.

In this section we investigate this question, focusing on the Novelty$^+$ algorithm. We made this choice because we observed SAPS's performance around $\langle \alpha, \rho\rangle = \langle 1.3, 0.1\rangle$ to be very close to optimal across many different instance distributions. SAPS thus offers little *possibility* for performance improvement through per-instance parameter tuning (Table 3 quantifies this), and so serves as a poor proving ground for our techniques. Novelty$^+$, on the other hand, exhibits substantial variation from one instance to another and from one instance distribution to another, making it a good algorithm for the evaluation of our approach.[6] We used the same test and training data as in the previous section; thus, Table 3 summarizes the experiments both from the previous section and from this section. However, in this section we also created a new instance distribution "Mixed", which is the union of the QWH and SAT04 distributions. This mix enables a large gain for automated parameter tuning (when compared to the best fixed parameter setting) since Novelty$^+$ performs best with high noise settings on unstructured instances and low settings on structured instances.

Figure 5(a) shows the performance of our automatic parameter-tuning algorithm on test data from Mixed, as compared to upper and lower bounds on its possible performance. We observe that the run-time with automatic parameter setting is very close to the optimal setting and far better than the worst one, with an increasing margin for harder instances. Figure 5(b) compares our automatic tuning against the best fixed parameter setting for the test set. This setting is often the most that can be hoped for in practice. (A common approach for tuning parameters is to perform a set of experiments,

[6] Indeed, the large potential gains for tuning WalkSAT's noise parameter on a per-instance basis have been exploited before [23].



(a) Automatic vs. best and worst

(b) Automatic vs. best fixed

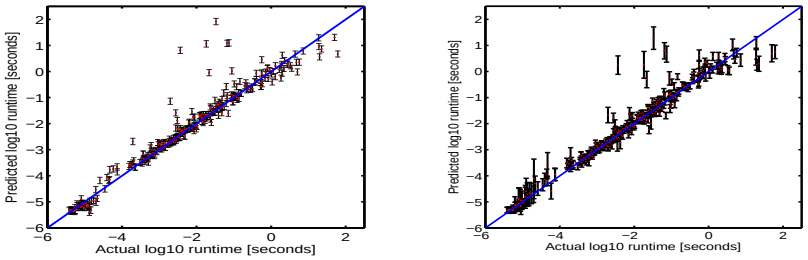
Fig. 5. (a) Performance of automated parameter setting for Novelty$^+$ on mixed data set QWH/SAT04, compared to the best (dots) and worst (crosses) per-instance parameter setting (out of the 6 parameter settings we employed). (b) Speedup of Novelty$^+$ over the best data-set specific fixed parameter setting.

to identify the parameter setting which achieves the lowest overall run-time, and then to fix the parameters to this setting.) Figure 5(b), in conjunction with Table 3 shows that our techniques dramatically outperform this form of parameter tuning. While Novelty$^+$ achieves an average speedup of over an order magnitude on Mixed as compared to the best fixed parameter setting on that set, SAPS improves upon its default setting by a factor of 2. Considering that our method is fully automatic and very general, these are very promising results.

4.1 Related work on automated parameter tuning

The task of configuring an algorithm's parameters for high and robust performance has been widely recognized as a tedious and time-consuming task that requires well-developed engineering skills. Automating this task is a very promising and active area of research. There exists a large number of approaches to find the best configuration for a given problem distribution [21, 3, 25, 1]. All these techniques aim to find a parameter setting that optimizes some scoring function which averages over all instances from the given input distribution. If the instances are very homogeneous, this approach can perform very well. However, if the problem instances to be solved come from heterogeneous distributions or even from completely unrelated application areas, the best parameter configuration may differ vastly from instance to instance. In such cases it is advisable to apply an approach like ours that can choose the best parameter setting for each run contingent on the characteristics of the current instance to be solved. This per-instance parameter tuning is more powerful but less general than tuning on a per-distribution basis in that it requires the existence of a set of discriminative instance features. However, we belief it to be relatively straight-forward to engineer a good set of instance features if one is familiar with the application domain.

The only other approach for parameter tuning on a per-instance basis we are aware of is the Auto-WalkSAT framework [23]. This approach is based on empirical findings showing that the optimal parameter setting of WalkSAT algorithms tends to be about 10% above the one that minimizes the invariance ratio [20]. Auto-WalkSAT chooses remarkably good noise settings on a variety of instances, but for domains where the



(a) Bayesian linear regression

(b) Gaussian process with squared exponential kernel

Fig. 6. Predictions and their uncertainty of Novelty$^+$ median run-time of 10 runs: trained on QCP, tested on QWH. The run-time predictions of these approaches are Gaussian probability distributions for every instance. The red dots specify the predictive mean and the black bars specify one standard deviation.

above relationship between invariance ratio and optimal noise setting does not hold (such as logistics problems), it performs poorly [23]. Furthermore, its approach is limited to SAT and in particular to tuning the (single) noise parameter of the WalkSAT framework. In contrast, our automated parameter tuning approach applies to arbitrary parametric algorithms and all domains for which good features can be engineered.

Finally, reactive search algorithms [2], such as Adaptive Novelty$^+$[13] or RSAPS [17] adaptively modify their search strategy *during* a search. (Complete reactive search algorithms include [15, 18, 7, 5].) Many reactive approaches still have one or more parameters whose settings remain fixed throughout the search; in these cases the automated configuration techniques we presented here should be applicable to tune these parameters. While a reactive approach is in principle more powerful than ours (it can utilize different search strategies in different parts of the space), it is also less general since the implementation is typically tighyly coupled to a specific algorithm. Ultimately, we aim to generalize our approach to allow for modifying parameters during the search — this requires that the features evaluated during search are very cheap. We also see reinforcement learning as very promising in this domain [18].

5 Uncertainty Estimates through Bayesian Regression

So far, research in empirical hardness models has focused on the case where the targeted application domain is known *a priori* and training instances from this domain are available. In practice, however, an algorithm may have to solve problem instances that are significantly different from the ones encountered during training. Empirical hardness models may perform poorly in this case. This is because the statistical foundations upon which their machine learning approach is built rely upon the test set being drawn from the same distribution as the training set. Bayesian approaches may be more appropriate in such scenarios since they explicitly model the *uncertainty* associated with their predictions. Roughly, they provide an automatic measure of how similar the basis functions for a particular test instance are to those for the training instances, and associate higher uncertainty with relatively dissimilar instances. We implemented two Bayesian methods: (a) sequential Bayesian linear regression (BLR) [4], a technique which yields

mean predictions equivalent to ridge regression but also offers estimates of uncertainty; and (b) Gaussian Process Regression (GPR) [24] with a squared exponential kernel. We detail BLR and the potential applications of a Bayesian approach to run-time prediction in an accompanying technical report [16]. Since GPR scales cubically in the number of data points, we trained it on a subset of 1000 data points (but used all 9601 data points for BLR). Even so, GPR took roughly 1000 times longer to train.

We evaluated both our methods on two different problems. The first problem is to train and validate on our QCP data-set and test on our QWH data-set. While these distributions are not identical, our intuition was that they share enough structure to allow models trained on one to make good predictions on the other. The second problem was much harder: we trained on data-set SAT04 and tested on a very diverse test set containing instances from ten qualitatively different distributions from SATLIB. Figure 6 shows predictions and their uncertainty ($\pm$ one stddev) for both methods on the first problem. The two distributions are similar enough to yield very good predictions for both approaches. While BLR was overconfident on this data set, the uncertainty estimates of GPR make more sense: they are very small for accurately predicted data points and large for mispredicted ones. Both models achieved similar predictive accuracy (CC/RMSE .97/.45 for BLR; .97/.43 for GPR). For the second problem (space considerations disallow a figure), BLR showed massive mispredictions (several tens of orders of magnitude) but associated very high uncertainty with the mispredicted instances, reflecting their dissimilarity with the training set. GPR showed more reasonable predictions, and also did a good job in claiming high uncertainty about instances for which predictive quality was low. Based on these preliminary results, we view Gaussian process regression as particularly promising and plan to study its application to run-time prediction in more detail. However, we note that its scaling behavior somewhat limits its applications.

6 Conclusion and Future Work

In this work, we have demonstrated that empirical hardness models obtained from linear basis function regression can be extended to make surprisingly accurate predictions of the run-time of randomized, incomplete algorithms, such as Novelty$^+$ and SAPS. Based on a prediction of sufficient statistics for run-time distributions (RTDs), we showed very good predictions of the entire empirical RTDs for unseen test instances. We have also demonstrated for the first time that empirical hardness models can model the effect of algorithm parameter settings on run-time, and that these models can be used as a basis for automated per-instance parameter tuning. In our experiments, this tuning never hurt and sometimes resulted in substantial and completely automatic performance improvements, as compared to default or optimized fixed parameter settings.

There are several natural ways in which this work can be extended. First, we are currently studying Bayesian methods for run-time prediction in more detail. We further plan to study the extent to which our results generalize to problems other than SAT and in particular, to optimization problems. Finally, we would like to apply active learning approaches [6] in order to probe the parameter space in the most informative way in order to reduce training time.

References

1. B. Adenso-Daz and M. Laguna. Fine-tuning of algorithms using fractional experimental design and local search. *Operations Research*, 54(1), 2006. To appear.
2. R. Battiti and M. Brunato. Reactive search: machine learning for memory-based heuristics. Technical Report DIT-05-058, Università Degli Studi Di Trento, Dept. of information and communication technology, Trento, Italy, September 2005.
3. M. Birattari, T. Stützle, L. Paquete, and K. Varrentrapp. A racing algorithm for configuring metaheuristics. In *Proc. of GECCO-02*, pages 11–18, 2002.
4. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.
5. T. Carchrae and J. C. Beck. Applying machine learning to low-knowledge control of optimization algorithms. *Computational Intelligence*, 21(4):372–387, 2005.
6. D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *JAIR*, 4:129–145, 1996.
7. S. L. Epstein, E. C. Freuder, R. J. Wallace, A. Morozov, and B. Samuels. The adaptive constraint engine. In *Proc. of CP-02*, pages 525 – 540, 2002.
8. C. Gebruers, B. Hnich, D. Bridge, and E. Freuder. Using CBR to select solution strategies in constraint programming. In *Proc. of ICCBR-05)*, pages 222–236, 2005.
9. I. P. Gent, H. H. Hoos, P. Prosser, and T. Walsh. Morphing: Combining structure and randomness. In *Proc. of AAAI-99*, pages 654–660, Orlando, Florida, 1999.
10. C. Gomes, B. Selman, N. Crato, and H. Kautz. Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *J. of Automated Reasoning*, 24(1), 2000.
11. C. P. Gomes and B. Selman. Problem structure in the presence of perturbations. In *Proc. of AAAI-97*, 1997.
12. H. H. Hoos. On the run-time behaviour of stochastic local search algorithms for SAT. In *Proc. of AAAI-99*, pages 661–666, 1999.
13. H. H. Hoos. An adaptive noise mechanism for WalkSAT. In *Proc. of AAAI-02*, pages 655–660, 2002.
14. H. H. Hoos and T. Stützle. *Stochastic Local Search - Foundations & Applications*. Morgan Kaufmann, SF, CA, USA, 2004.
15. E. Horvitz, Y. Ruan, C. P. Gomes, H. Kautz, B. Selman, and D. M. Chickering. A Bayesian approach to tackling hard computational problems. In *Proc. of UAI-01*, 2001.
16. F. Hutter and Y. Hamadi. Parameter adjustment based on performance prediction: Towards an instance-aware problem solver. Technical Report MSR-TR-2005-125, Microsoft Research, Cambridge, UK, December 2005.
17. F. Hutter, D. A. D. Tompkins, and H. H. Hoos. Scaling and probabilistic smoothing: Efficient dynamic local search for SAT. In *Proc. of CP-02*, volume 2470, pages 233–248, 2002.
18. M. G. Lagoudakis and M. L. Littman. Learning to select branching rules in the DPLL procedure for satisfiability. In *Electronic Notes in Discrete Mathematics (ENDM)*, 2001.
19. K. Leyton-Brown, E. Nudelman, and Y. Shoham. Learning the empirical hardness of optimization problems: The case of combinatorial auctions. In *Proc. of CP-02*, 2002.
20. D. McAllester, B. Selman, and H. Kautz. Evidence for invariants in local search. In *Proc. of AAAI-97*, pages 321–326, 1997.
21. S. Minton. Automatically configuring constraint satisfaction programs: A case study. *Constraints*, 1(1):1–40, 1996.
22. E. Nudelman, K. Leyton-Brown, H. H. Hoos, A. Devkar, and Y. Shoham. Understanding random SAT: Beyond the clauses-to-variables ratio. In *Proc. of CP-04*, 2004.
23. D. J. Patterson and H. Kautz. Auto-WalkSAT: a self-tuning implementation of walksat. In *Electronic Notes in Discrete Mathematics (ENDM), 9*, 2001.
24. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
25. B. Srivastava and A. Mediratta. Domain-dependent parameter selection of search-based algorithms compatible with user performance criteria. In *Proc. of AAAI-05*, 2005.