Statistical Generalization Of Performance-Related Heuristics for Knowledge-Lean Applications *

Arthur Ieumwananonthachai and Benjamin W. Wah

Center for Reliable and High Performance Computing
Coordinated Science Laboratory
University of Illinois, Urbana-Champaign
1308 West Main Street, Urbana, IL 61801, USA
{arthuri, wah}@manip.crhc.uiuc.edu

Abstract

In this paper, we present new results on the automated generalization of performance-related heuristics learned for knowledge-lean applications. We study methods to statistically generalize new heuristics learned for some small subsets of a problem space (using methods such as genetics-based learning) to unlearned problem subdomains. Our method uses a new statistical metric called probability of win. By assessing the performance of heuristics in a range-independent and distribution-independent manner, we can compare heuristics across problem subdomains in a consistent manner. To illustrate our approach, we show experimental results on generalizing heuristics learned for sequential circuit testing, VLSI cell placement and routing, and branch-and-bound search. We show that generalization can lead to new and robust heuristics that perform better than the original heuristics across problem instances of different characteristics.

1 Introduction

Heuristics or heuristic methods (HMs), in general terms, are "Strategies using readily accessible though loosely applicable information to control problem-solving processes in human being and machines" [1]. They exist as problem solving procedures in problem solvers to find (usually) suboptimal solutions for many engineering applications. Since their design depends on user experience and is rather ad hoc, it is desirable to acquire them automatically by machine learning.

In this paper, we focus on applications that are *knowledge-lean*, implying that domain knowledge for credit assignment is missing. In this class of applications, we are interested to learn and generalize *performance-related* HMs whose goal is to find solutions with the best numerical performance. Examples of targeted HMs and applications include symbolic formula for guiding decision making process in a branch-and-bound search and a set of numerical parameters used in a simulated annealing package for VLSI circuit placement and routing. (see Section 4).

Genetics-based learning is one of the methods developed for learning and generalization for the applications studied in this paper [2 3]. This is a form of learning by induction that involves applying genetic algorithms [4] to machine learning problems. There are two steps involved in this learning method:

(a) *generation* and *selection* of HMs that can better solve test cases used in learning, as compared to the best existing (baseline) HMs;

(b) *generalization* of the selected HMs to test cases not seen in learning with the same high level of performance as compared to that of the baseline HMs.

As illustrated in Figure 1, these two steps are generally separated in genetics-based learning.

In this paper, we study *statistical generalization of HMs across test cases of an application with different performance distributions.* When the performance of a HM across different test cases is of different distributions, statistical metrics like average cannot be applied. Our approach in this paper is to (a) partition the domain of test cases into *subdomains* in such a way that performance values in a subdomain are *independent and identically distributed* (i.i.d.), and (b) develop conditions under when a HM can be considered to perform well across multiple subdomains. Note that we do not modify a HM in order to generalize it across subdomains (as studied in artificial intelligence [5]) but rather

*Research supported by National Science Foundation Grant MIP 92-18715 and National Aeronautics and Space Administration Grant NAG 1-613.

IEEE International Conference on Tools with Artificial Intelligence, November 1995.

Table 1: Maximum and average fault coverages of two HMs used in a test-pattern generator with different random seeds.

Circuit	HM	Maximum FC	Average FC
S444	101	60.3	28.5
	535	86.3	84.8
S1196	101	94.9	94.2
	535	93.6	93.1

cult because test cases in different subdomains of a domain may have different performance distributions, even though they can be evaluated by a common HM. As a result, the performance of test cases cannot be compared statistically.

As an example, Table 1 shows the average and maximum fault coverages of two HMs used in a test-pattern generator to test sequential circuits. The data indicate that we cannot average their fault coverages across the two circuits as the performance distribution of HM_{101} across the two circuits is not the same as that of HM_{535}.

It should now be clear that there can be many subdomains in an application, and learning can only be performed on a small number of them. Consequently, it is important to generalize HMs learned for a small number of subdomains to other subdomains. In some situations, multiple HMs may have to be identified and applied together at a higher cost to find a solution of higher quality.

3 Generalization of Heuristic Methods Learned

Given the definition of a problem subdomain, we can now identify two issues in generalizing HMs across test cases. First, we need to evaluate and generalize the performance of a HM in a single subdomain in a range-independent and distribution-independent way. Only then can we address the issue on performance evaluation and generalization across multiple subdomains.

3.1 Performance Evaluation within a Subdomain

There are many ways to address the first issue raised above, and solutions to the second issue depend on the solution to the first. For instance, scaling and normalization of performance values is a possible way to compare performance in a distribution-independent manner; however, this may lead to new inconsistencies [2]. Another way is to rank HMs by their performance values and use the average ranks of HMs for comparison. This does not work well because it does not account for actual differences in performance values, and two HMs with very close or very different performance may differ only by one in their ranks. Further, the maximum rank of HMs depends on the number of HMs evaluated, thereby biasing the average ranks of individual HMs. In this section, we propose a metric called probability of win to select good HMs within a subdomain.

$P_{win}(h_i, d_m)$, the *probability-of-win* of HM h_i in subdomain d_m, is defined as the probability that the true mean of h_i (on one performance measure[1]) is better than the true mean of HM h_j randomly selected from the pool. When h_i is applied on test cases in d_m, we have

$$P_{win}(h_i, d_m) = \frac{\sum_{j \neq i} P\left[\mu_i^m > \mu_j^m \,|\, \hat{\mu}_i^m, \hat{\sigma}_i^m, n_i^m, \hat{\mu}_j^m, \hat{\sigma}_j^m, n_j^m\right]}{|s| - 1}, \quad (1)$$

where $|s|$ is the number of HMs under consideration, and n_i^m, $\hat{\sigma}_i^m$, $\hat{\mu}_i^m$, and μ_i^m are, respectively, the number of tests, sample standard deviation, sample mean, and true mean of h_i in d_m.

Since we are using the average performance metric, it is a good approximation to use the normal distribution as a distribution of the sample average. The probability that h_i is better than h_j in d_m can now be computed as follows.

$$P\left[\mu_i^m > \mu_j^m \,|\, \hat{\mu}_i^m, \hat{\sigma}_i^m, n_i^m, \hat{\mu}_j^m, \hat{\sigma}_j^m, n_j^m\right] \approx \Phi\left[\frac{\hat{\mu}_i^m - \hat{\mu}_j^m}{\sqrt{\hat{\sigma}_i^{m2}/n_i^m + \hat{\sigma}_j^{m2}/n_j^m}}\right] \quad (2)$$

where $\Phi(x)$ is the cumulative distribution function for the $N(0,1)$ distribution.

To illustrate the concept, we show in Table 2 the probabilities of win of four HMs tested to various degrees. Note that P_{win} is not only related to the sample mean but also depends on the sample variance and number of tests performed. Further, the probability that h_i is better than h_j and the probability that h_j is better than h_i are both counted in the evaluation. Hence, the average of P_{win} over all HMs in a subdomain ($= \sum_i P_{win}(h_i, d_m)/|s|$) will be 0.5.

P_{win} defined in Eq. (1) is range-independent and distribution-independent because all performance values are transformed into probabilities between 0 and 1 independent of the number of HMs evaluated and the distribution of performance values. It assumes that all HMs are i.i.d. and takes into account uncertainty in their sample averages (by using their variances); hence, it is better than simple scaling that only compresses performance averages into a range between 0 and 1. It is also important to point out that the HMs used

[1] Due to space limitation, we do not consider issues dealing with multiple performance measures in this paper.

Table 2: Probabilities of win of four HMs in d_m.

h_i	$\hat{\mu}_i$	$\hat{\sigma}_i$	n_i	$P_{win}(h_i, d_m)$
1	43.2	13.5	10	0.4787
2	46.2	6.4	12	0.7976
3	44.9	2.5	10	0.6006
4	33.6	25.9	8	0.1231

in computing P_{win} are found by learning; hence, they already perform well within a subdomain.

3.2 Performance Evaluation across Subdomains

One of the major difficulties in handling multiple subdomains is that it may be difficult to aggregate performance values statistically from different subdomains, and to define the notion that one HM performs better than another across multiple subdomains. For instance, it is not meaningful to find an average of random numbers from two different distributions. We address this problem using P_{win} defined in the last subsection.

First, we assume that when HM h is applied over multiple subdomains in partition Π_p of subdomains, all subdomain are equally likely. Therefore, we compute P_{win} of h over subdomains in Π_p as the average P_{win} of h over all subdomains in Π_p.

$$P_{win}(h, \Pi_p) = \frac{\sum_{d \in \Pi_p} P_{win}(h, d)}{|\Pi_p|}, \qquad (3)$$

where Π_p is the p'th partition of subdomains in the problem domain. The HM picked is the one that maximizes Eq. (3). When subdomains are not equally likely but with known relative weights, we can compute P_{win} as a weighted average instead of Eq. (3). HMs picked using Eq. 3 generally wins with a high probability across most of the subdomains in Π_p but occasionally may not perform well in a few subdomains.

Second, we consider the problem of finding a good HM across multiple subdomains in Π_p as a multi-objective optimization problem. In this case, evaluating HMs based on a combined objective function (such as the average P_{win} in Eq. (3) may lead to inconsistent conclusions. To alleviate such inconsistencies, we should treat each subdomain independently and find a common HM across all subdomains in Π_p satisfying some common constraints. For example, let δ be the allowable deviation of P_{win} of any chosen HM from q^m_{win}, the maximum P_{win} in subdomain m. Generalization, therefore, amounts to finding h that satisfies the following constraints for every subdomain $m \in \Pi_p$.

$$P_{win}(h, m) \geq (q^m_{win} - \delta) \qquad \forall m \in \Pi_p \qquad (4)$$

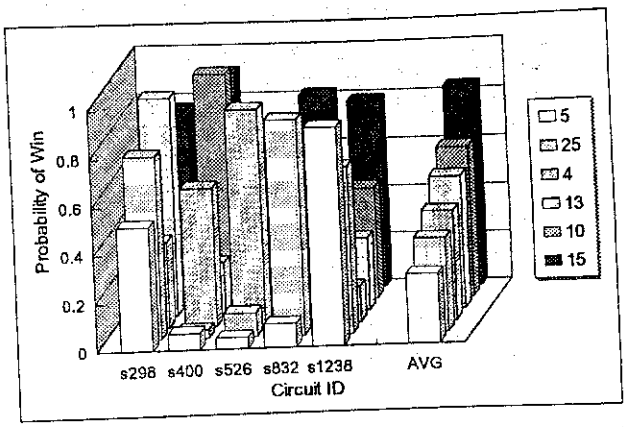

Figure 2: P_{win} of six HMs across five subdomains in the test-pattern generation problem.

Here, δ may need to refined if there are too many or too few HMs satisfying the constraints.

To illustrate the generalization procedure, consider the test-pattern generation problem discussed in Section 2. Assume that learning had been performed on five circuits (subdomains), and that the six best HMs from each subdomain were reported. After full evaluation of the 30 HMs (initialized by ten random seeds) across all five subdomains, we computed P_{win} of each HM in every subdomain. Figure 2 shows the probabilities of win of six of these HMs. If we generalize HMs based on Eq. (3), then HM_{15} will be picked since it has the highest average P_{win}. Likewise, if we generalize using Eq. (4), we will also select HM_{15}. Note that in this example, no one HM is the best across all subdomains.

4 Experimental Results

To illustrate the generalization procedure described in Section 3, we present in this section results on generalization for two applications in VLSI design and branch-and-bound search. These results were obtained using TEACHER [2], a genetics-based learning system that implements our proposed generalization strategy.

4.1 HM for Sequential Circuit Testing

The first application is based on CRIS [8], a genetic-algorithm software package for generating patterns to test sequential VLSI circuits. CRIS mutates an input test sequence continuously and analyzes the mutated vectors in selecting a test set. Since many copies of a circuit may be manufactured, it is desirable to obtain as high a fault coverage as possible, and computational cost is of secondary importance.

In our experiments, we used sequential circuits from the ISCAS89 benchmarks [9] plus several other larger circuits. We treat each circuit as an individual subdomain. Since we want one common HM for all circuits,

Table 3: Parameters in CRIS treated as a HM in learning and in generalization. (The type, range, and step of each parameter were given to us by the designer of CRIS. The default parameters were not given to us as they are circuit-dependent.)

Par.	Range	Step	Definition	New Value
P_1	1-10	1	related to the number of stages in a flip flop	1
P_2	1-40	1	sensitivity of state change of a flip flop	12
P_3	1-40	1	survival rate of a test sequence in next generation	38
P_4	0.1-10.0	0.1	number of test vec. concat. to form a new vec.	7.06
P_5	50-800	10	number of useless trials before quitting	623
P_6	1-20	1	number of generations	1
P_7	0.1-1.0	0.1	how genes are spliced in GA	0.1
P_8	Integer	1	seed for random number generator	-

we assume that all circuits are from one domain.

CRIS in our experiments is treated as a black-box problem solver, as we have minimal knowledge in its design. A HM targeted for improvement is a set of eight parameters used in CRIS (Table 3). Note that parameter P_8 is a random seed, implying that CRIS can be run multiple times using different random seeds in order to obtain better fault coverages. (In our experiments, we used a fixed sequence of ten random seeds.)

In our experiments on CRIS, we chose five circuits as our learning subdomains. In each of these subdomains, we used TEACHER [2], a genetics-based learning system, to test CRIS 1000 times (divided into 10 generations) with different HMs. A HM in learning is represented as a tuple of the first seven parameters in Table 3. At the end of learning, we picked the top twenty HMs in each subdomain and evaluated them fully by initializing CRIS using ten different random seeds (P_8 in Table 3). We then selected the top five HMs from each subdomain, resulting in a total of 25 HMs supplied to the generalization phase. We evaluated the 25 HMs fully (each with 10 random seeds) on the five subdomains used in learning and five new subdomains. We then selected one generalized HM to be used across all the ten circuits (based on Eq. (3)). The HM found is shown in the last column in Table 3.

Table 4 summarizes the improvements of our learned and generalized HMs as compared to the published results of CRIS [8] and HITEC [10]. Each entry of the table shows the number of times our HM wins and ties in terms of fault coverage with respect to the method(s)

Table 4: Summary of results comparing the per... mance of our generalized HMs with respect to ... of HITEC and CRIS. (The first number in each ... shows the number of wins out of 21 circuits, and ... second, the number of ties.)

Our HM wins/ties with respect to the following	CRIS Generalized HM	
	Max. FC	Avg. FC
HITEC	6, 1	4, 0
CRIS	16, 1	11, 0
Both HITEC and CRIS	5, 2	3, 0

in the first column. Note that the maximum fault co... erages reported in Table 4 were based on ten runs ... the underlying problem solver, implying that the com... putational cost is ten times of the average cost. Recall that we like to obtain the maximum coverage of a cir... cuit, and that computational cost is a secondary issue in circuit testing.

Our results show that our generalization procedure can discover new HMs that are better than the original HMs in 16 out of 21 circuits in terms of the maximum fault coverage, and in 11 out of 21 circuits in terms of the average fault coverage. Our results are significant in the following aspects:

(a) new faults detected by our generalized HMs were not discovered by previous methods;

(b) only one HM (rather than many circuit-dependent HMs in the original CRIS) was found for all circuits.

Table 4 also indicates that HITEC is still better than our new generalized HM for CRIS in most of the circuits. This happens because our generalized HM is bounded by the limitations in CRIS and our HM generator for CRIS. Such limitations cannot be overcome without generating more powerful HMs in our HM generator or using better test-pattern generators like HITEC as our baseline problem solver.

4.2 HM for VLSI Placement and Routing

In our second application, we use TimberWolf [11] as our problem solver. This is a software package based on simulated annealing (SA) [12] to place and route various components on a piece of silicon. Its goal is to minimize the chip area needed while satisfying constraints such as the number of layers of poly-silicon for routing and the maximum signal delay through any path.

Although in theory SA converges asymptotically to the global optimum with probability one, the results generated in finite time are usually suboptimal. As

Table 5: Parameters in TimberWolf (Version 6) used in our HM for learning and for generalization.

Par.	Range	Step	Meaning	Orig.	New
P_1	0.1 - 2.5	0.1	vertical path weight for estimating the cost function	1.0	0.958
P_2	0.1 - 2.5	0.1	vertical wire weight for estimating the cost function	1.0	0.232
P_3	3 - 10	1	orientation ratio	6	10
P_4	0.33 - 2.0	0.1	range limiter window change ratio	1.0	1.30
P_5	10.0 - 35.0	1.0	high temperature finishing point	23.0	10.04
P_6	50.0 - 99.0	1.0	intermediate temperature finishing point	81.0	63.70
P_7	100.0 - 150.0	1.0	low temperature finishing point	125.0	125.55
P_8	130.0 - 180.0	1.0	final iteration temperature	155.0	147.99
P_9	0.29 - 0.59	0.01	critical ratio that determines acceptance probability	0.44	0.333
P_{10}	0.01 - 0.12	0.01	temperature for controller turn off	0.06	0.112
P_{11}	integer	1	seed for the random number generator	-	-

a result, there is a trade-off between the quality of a result and the cost (or computational time) of obtaining it. In TimberWolf version 6.0, the version we have studied, there are two parameters to control the running time (which indirectly control the quality of the result): *fast-n* and *slow-n*. The larger the *fast-n* is, the shorter time SA will run. In contrast, the larger the *slow-n* is, the longer time SA will run. Of course, only one of these parameters can be used at any time.

TimberWolf has six major components: *cost function*, *generate function*, *initial temperature*, *temperature decrement*, *equilibrium condition*, and *stopping criterion*. In Table 5, we list the parameters we have focused in this study. Our goal is to illustrate the power of our learning and generalization procedures and to show improved quality and reduced cost for the placement and routing of large circuits, despite the fact that only small circuits were used in learning.

In our experiments, we used seven benchmark circuits [13] (s298, s420, fract, primary1, struct, primary2, industrial1). We studied only the standard-cell placement problem, noting that other kinds of placement can be studied in a similar fashion. We used *fast-n* values of 1, 5, and 10, respectively.

We first applied TEACHER to learn good HMs for

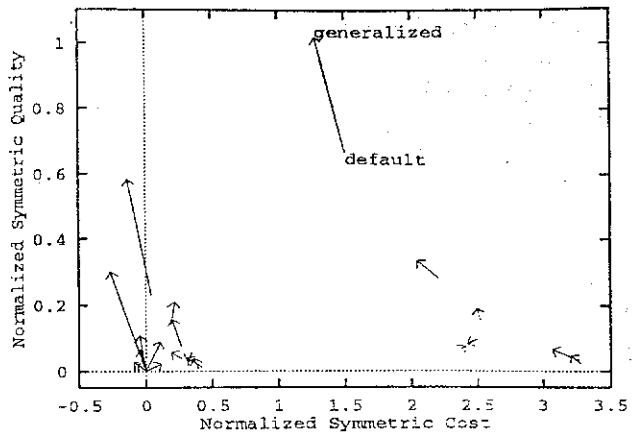

Figure 3: Comparison of normalized average performance between the default and the generalized HMs. The plots are normalized with respect to the performance of applying the baseline HM on each circuit using *fast-n* = 10. (See Eq. (5)).

circuits *s298* with *fast-n* of 1, *s420* with *fast-n* of 5, and *primary1* with *fast-n* of 10, each of which was taken as a learning subdomain. We used a fixed sequence of ten random seeds (P_{11} in Table 5) in each subdomain to find the statistical performance of a HM. Each learning experiment involved 1000 applications of TimberWolf divided into ten generations. Based on the best 30 HMs (10 from each subdomain), we applied our generalization procedure to obtain one generalized HM. This generalized HM as well as the default HM are shown in Table 5.

Figure 3 plots the quality (higher quality in the y-axis means reduced chip area averaged over 10 runs using the defined random seeds) and cost (average execution time of TimberWolf) between the generalized HM and the default HM on all seven circuits with *fast-n* of 1, 5, and 10, respectively. Note that all performance values in Figure 3 are normalized with respect to those of *fast-n* of 10, and that the positive (resp., negative) portion of the x-axes shows the fractional improvement (resp., degradation) in computational cost with respect to the baseline HM using *fast-n* of 10 for the same circuit. Each arrow in this figure points from the average performance of the default HM to the average performance of the generalized HM.

The equation for computing the normalized symmetric cost is as follows. Let C_{new}, C_{base} and C_{sym}^{norm} be, respectively, the costs of the new HM, the cost of the baseline HM, and the normalized symmetric cost.

$$C_{sym}^{norm} = \begin{cases} \frac{C_{new}}{C_{base}} - 1 & \text{if } C_{new} \geq C_{base} \\ 1 - \frac{C_{base}}{C_{new}} & \text{if } C_{new} < C_{base} \end{cases} \quad (5)$$

The reason for using the above equation is to avoid uneven compression of the ratio C_{new}/C_{base}. This ratio is between 0 and 1 when $C_{new} < C_{base}$, but is between 1 and ∞ when $C_{new} > C_{base}$. Eq. (5) allows increases in cost to be normalized in the range between 0 and ∞, and decreases to be normalized in the range between 0 and $-\infty$. The normalized symmetric quality in the y-axis is computed in a similar way.

Among the 21 test cases, the generalized HM has worse quality than that of the default in only two instances, and has worse cost in 4 out of 21 cases. We see in Figure 3 that most of the arrows point in a left-upward direction, implying improved quality and reduced cost. We expect to see more improvement as we learn other functions and parameters in TimberWolf. Further, improvements in TimberWolf are important as the system is actually used in industry.

4.3 Branch-and-Bound Search

A branch-and-bound search algorithm is a systematic method for traversing a search tree or search graph in order to find a solution that optimizes a given objective while satisfying the given constraints. It decomposes a problem into smaller subproblems and repeatedly decomposes them until a solution is found or infeasibility is proved. Each subproblem is represented by a node in the search tree/graph. In this subsection, we apply learning to find new *decomposition HMs* for expanding a search node into descendants.

We illustrate our method on three applications: traveling salesman problem (TSP) on incompletely connected graphs mapped on a two-dimensional plane, vertex-cover problem (VC), and knapsack problem (KS). We assume that each problem constitutes one domain.

We use well-known decomposition HMs developed for these applications as our baseline HMs (see Table 6). The normalized cost of a candidate decomposition HM is defined in terms of its *average symmetric speedup* (see Eq. (5)), which is related to the number of nodes expanded by a branch-and-bound search using the baseline HM and that using the new HM. Note that we do not need to measure quality as both the new and existing HMs when applied in a branch-and-bound search look for the optimal solution.

In our experiments, we selected six subdomains in each application for learning. We performed learning in each subdomain using 1,600 tests, selected the top five HMs in each subdomain, fully verified them on all the learned subdomains, and selected one final HM to be used across all the subdomains (Eq. (3)). Table 7 summarizes the generalization and validation results. Our results show that we have between 0-8% improvement in average symmetric speedups using the generalized HMs. Note that the baseline HM is the best HM for solving the knapsack problem.

Table 6: Original and generalized decomposition HMs used in a branch-and-bound search (l: number of uncovered edges or live degree of a vertex; n: average live degree of all neighbors; Δl: difference between l of parent node and l of current node; c: length of current partial tour; m: minimum length to complete current tour; p: profit of object; w: weight of object).

Application	Original HM	Generalized HM
VC	l	$1000\,l + n - \Delta l$
TSP	c	$m\,c$
KS	p/w	p/w

The second part of Table 7 shows the average symmetric speedups when we validate the generalized HMs on larger test cases. These test cases generally require 10-50 times more nodes expanded than those used earlier. Surprisingly, our results show better improvement (9-23%). It is interesting to point out that six of the twelve subdomains with high degree of connectivity in the vertex-cover problem have slowdowns. This is a clear indication that these subdomains should be grouped in a different domain and learned separately.

Table 6 shows the new decomposition HMs learned for the three applications. An example is the HM learned for the vertex cover problem. This formula can be interpreted as using l as the primary key for deciding which node to include in the covered set. If the l's of two alternatives are different, then the remaining terms in the formula $(n - \Delta l)$ are insignificant. On the other hand, when the l's are the same, then we use $(n - \Delta l)$ as a tie breaker.

5 Conclusions

In this paper, we have presented a method for generalizing performance-related heuristics learned by genetics-based learning for knowledge-lean applications. We have focused on a class of heuristic methods (HMs) whose performance is evaluated statistically by applying them on multiple test cases. Due to a lack of domain-knowledge for improving such heuristics, we have used a genetics-based learning paradigm (a generate-and-test method) to learn new HMs.

We have proposed in this paper a new metric called probability of win to characterize the performance of heuristics. This metric evaluates the probability that the mean performance of a HM is better than the mean performance of another randomly chosen HM in a set of learned HMs on a common set of test cases. The only requirement on the choice of test cases in evaluating probabilities of win is that each HM, when evaluated on the test cases, produces a set of independent and identically distributed performance results. We define