Optimizations for the Boolean Approach to Computing Minimal Hitting Sets

Ingo Pill and Thomas Quaritsch

Abstract. The Boolean approach to computing minimal hitting sets proposed by Lin and Jiang is known to offer very attractive general performance, but also has its issues, specifically with a cardinality-restricted search. In this paper we propose optimizations regarding the refinement rules, also offering a revised decision strategy as well as optimized termination criteria that exploit cardinality bounds. Our experiments including artificial and real-world samples for the bounded and unbounded case show the potential of our work, where we could achieve speed-ups of up to two orders of magnitude.

1 INTRODUCTION

When describing a problem as a set of structures, minimal hitting sets that contain at least one element per structure often encode important aspects, and may even provide the solutions to essential problems. For instance, in model-based diagnosis, the minimal hitting sets for a set of unsatisfiable cores (conflict sets) provide the diagnoses [4, 9, 5, 11, 1], where some more recent ones [10, 3] may trade perturbations entailed the development of many competing approaches, e.g. may draw on minimal hitting sets. The widespread application of program slices are used to identify program faults [12]. Even such low-level tasks as transforming conjunctive and disjunctive normal forms may draw on minimal hitting sets. The widespread application options entailed the development of many competing approaches, e.g. [9, 5, 4, 11, 1], where some more recent ones [10, 3] may trade performance with completeness. That is, aiming at some minimal solution, they might shed others during optimization stages, as long as the derived one is in fact a minimal one. For some applications, one’s attention might be however on complete approaches, that is those that can deliver the entire set of subset-minimal hitting sets. Lin and Jiang proposed in [7] such an approach considering the sets to hit in a structured way, and showed how to efficiently implement their idea using bits for components. This Boolean approach is well-known to offer very good performance, but also has its limitations. That is, if we establish cardinality limits on the the desired solutions (due to resource limits, or if too complicated solutions are not of interest), the Boolean approach, unlike the approaches of Reiter [9, 5] (HS-DAG) and Wotawa [11] (HST), might struggle with exploiting these search space limitations. In this paper we focus on such issues and propose optimizations that help to tackle them.

We organized our paper as follows. We give a formal introduction of the minimal hitting set problem and the Boolean approach in Section 2. Section 3 offers an analysis of the issues at hand and provides the details of our optimizations. In Section 4, we discuss our test scenarios, with Section 4.1 focusing on the test platform and our results. Finally, we draw our conclusions in Section 5.

2 THE MINIMAL HITTING SET PROBLEM AND THE BOOLEAN ALGORITHM

According to Reiter [9], a hitting set and a minimal hitting set are formally defined as follows:

Definition 1 Given a set of sets SCS, a set h ⊆ ∪i∈SCS CSi ⊆ COMP is a hitting set for SCS, iff for any set CSi ∈ SCS the intersection with h is non-empty, i.e. h ∩ CSi ≠ Ø.

Definition 2 Given a hitting set h for some SCS, h is said to be minimal with respect to subset inclusion, iff there exists no other hitting set h′ for SCS, such that h′ ⊂ h.

While from here on we assume subset-minimality when referring to minimal hitting sets (MHSs) there are also other attractive dimensions for minimality. Let us consider the following example.

Example 1 Let SCS be the set \{\{1, 2, 5\}, \{2, 4\}, \{2, 3\}\}. Then, the sets \{2\}, \{1, 4, 3\}, and \{5, 4, 3\} are (subset-)minimal hitting sets for SCS, while obviously the first one is much smaller than the latter two.

Thus, when an MHS’s cardinality correlates with a problem’s solution’s complexity, one might actually consider cardinality restrictions on the search space, specifically as some algorithms such as HS-DAG [9, 5] and HST [11] tend to profit greatly from such restrictions via the resulting depth limits on their internal trees. Sometimes also limits on computation resources require us to focus only on the most attractive (in the current context smallest) solutions.

Another option in the same direction is to exploit probabilities (weights). That is, in the context of model-based diagnosis, de Kleer and Williams define the probability of an MHS (a diagnosis Δ) [4] as \(\prod_{c∈Δ} p_f(x) \cdot \prod_{c∈COMP\setminusΔ} (1 - p_f(x))\), where COMP is the set of components, and \(p_f\) defines the probability of \(c ∈ COMP\)’s appearance in \(Δ\) while assuming stochastic independence between the components’ involvement. While in this paper we take cardinality limits into account, our findings can be extended to consider other metrics like probabilities.

As mentioned earlier, in 2003 Lin and Jiang proposed the Boolean approach [7] using bits (propositions) for components \(c \in COMP\) in order to derive all minimal hitting sets for a given SCS. They encode SCS as a Boolean formula in disjunctive normal form (DNF), with the conjuncts encoding the individual \(CS_i \in SCS\) and consisting of the corresponding (negated) element bits. A recursive function \(H(C)\) containing five rules (considered in ascending order) derives from this SCS formula another formula encoding the MHSs. That is, the result still needs some subset-checks (or the use of Boolean laws) in order to derive a canonical DNF where the conjuncts represent the individual MHSs. Assuming \(C\) a Boolean formula, \(e\) an atomic

1 Institute for Software Technology, Graz University of Technology, Infelfsgasse 16b2, 8010 Graz, Austria, email: {pill, quaritsch}@ist.tugraz.at
prophecy with \(\bar{e}\) denoting its negation, and False/True referring to
\(e \wedge \bar{e} \vee \bar{e}, H(C)\) is defined as:

R1: \(H(\text{False}) = \text{True}, \ H(\text{True}) = \text{False};\)
R2: \(H(\bar{e}) = e;\)
R3: \(H(\bar{e} \wedge C) = e \lor H(C);\)
R4: \(H(\bar{e} \lor C) = e \lor H(C);\)
R5: \(H(C) = \bar{e} \lor H(C_1) \lor H(C_2)\) for some arbitrary atomic
proposition \(e\) present in \(C\), with \(C_1 = \{c | e \in C \land \bar{e} \not\in c\}\)
and \(C_2 = \{c | e \not\in c \land (c \in C \lor \bar{e} \in c)\}\).

Please note that like Lin and Jiang we consider a conjunction also
as a set of elements. R5 encodes the algorithm’s general strategy of
how to conquer the search space. That is, splitting on some proposition
\(e \in \text{COMP}\), the algorithm forks two branches; the “left” one that
considers those solutions containing \(e\) (and thus subsequently focusing
on those \(CS_i\)’s not hit so far), while the “right” one assumes that \(e\)
is not part of the solution (with a further focus on all sets, but with \(\bar{e}\)
removed from the problem description – \(C\) is replaced by \(C_2\)). R1 to
R4 resolve specific situations, that is, R4 covers the situation when
there is a \(CS_i\) with only one element (there is one obvious choice
then), R3 and R2 those situations where \(|SCS| = 1\), and R1 resolves
True and False.

Obviously, the decision heuristic choosing the split element \(e\) in
R5 has a significant impact on the actual traversal of the search
space. A common heuristic (which we will refer to as H1) is to
consider those solutions containing \(e\) and thus subsequently focusing
on those \(CS_i\)’s not hit so far, while the “right” one assumes that \(e\)
is not part of the solution (with a further focus on all sets, but with \(\bar{e}\)
removed from the problem description – \(C\) is replaced by \(C_2\)). R1 to
R4 resolve specific situations, that is, R4 covers the situation when
there is a \(CS_i\) with only one element (there is one obvious choice
then), R3 and R2 those situations where \(|SCS| = 1\), and R1 resolves
True and False.

Like for H1, the validity of H2 follows from the fact that as
allowed by the original paper we could have chosen any \(e \in \text{elements}(C)\), and we choose some \(e \in CS_i \subseteq \text{elements}(C)\).

Intuitively, H2 strives to “clear” some (minimal) \(CS_i\) as fast as
possible, where for the last element in \(CS_i\), R4 takes over. When
discussing the options for cardinality-limit related breaks in the
algorithm later on, it will become clear that this effectively limits the
amount of right branches to be considered. If there are multiple sets
of minimal length, we might encounter some (non-harming)
non-determinism resulting in “hopping” between those sets. For any
recursive call of \(H(C)\) removing another element from a minimal \(CS_i\),
we also have the checks whether R1 to R4 would apply (even when
\(|CS_i| > 1\) and \(|SCS| > 1\), which requires R5). Thus we propose the
following variant of R5 that loops on a single \(CS_i\) directly.

Lemma 2 Adapting R5 as follows does not affect the algorithm’s
correctness:
\(H(C \lor \bigwedge_{i=1}^{n} \bar{e}_i) = \bigvee_{i=1}^{n-1} (e_i \land H(C_1^i)) \lor H(C^{i-1})\)
where \(C^0 = C, \forall c_i \in C: |c_i| \geq n, \text{ and the sets } C^i \text{ and } C_1^i \text{ are}
defined as
\(C^i = \{c_j | e_i \not\in c_j \land (c_j \in C^{i-1} \lor e_i \in C^{i-1})\}\)
\(C_1^i = \{c_j | c_j \in C^{i-1} \land e_i \not\in c_j\}.

Proof: An essential pre-condition for the correctness is our focus on
some cardinality-minimal \(CS_i\) as ensured by the prerequisites. Thus,
when removing the \(n - 1\) elements from \(CS_i\), in the sequence of the sets \(C_1^i\), there exists no other \(CS_i\) such that its refinements could
trigger R1 to R4. Intuitively, and implementing the general split mecha-
nism of R5, the disjunction of terms \(e_i \land H(C_i^1)\) offers those terms
established by the left branches of each recursion of the original R5,
when choosing the same sequence of elements in \(CS_i\) for H2 as split
elements (the right branch acting as interface between the recursive
calls). \(H(C^{i-1})\) is the last missing right branch, i.e. the one that
triggers R4 in order to deal with the n-th element in \(CS_i\).
The following example illustrates Lemma 2, where we abbreviate \( b \land c \land g \) with \( b \bar{c} g \), and loop on the underlined \( CS \).

**Example 2** \( H(C^b) = H(b \bar{c} g \lor a \bar{f} g \lor a \bar{c} \bar{d} \lor b \bar{c} f) = b H(C_1^b) \lor \neg b H(C_2^b) \lor \neg b H(C_3^b) \); where
\[
C_1^b = b \bar{c} g \lor a \bar{f} g \lor a \bar{c} \bar{d} \lor b \bar{c} f, \quad C_2^b = \neg b \lor b \bar{c} g \lor a \bar{f} g \lor a \bar{c} \bar{d} \lor b \bar{c} f, \quad C_3^b = \neg b \lor b \bar{c} g \lor a \bar{f} g \lor a \bar{c} \bar{d} \lor b \bar{c} f.
\]

R5' makes the connection between our strategy as implemented in both H2 and R5' with the tree-construction used in HS-DAG most obvious, as we capture all the essential details that allow HS-DAG to cut the depth of the internal tree. Thus, specifically for scenarios with \( |MHS| = 1 \), we expect a huge performance benefit.

Other options for performance gains are unveiled when focusing on restriction-related termination criteria in the algorithm. That is, in contrast to simply discarding larger solutions, the computation will focus only on branches that can produce viable MHSs. Explicitly keeping track of an intermediate solution’s cardinality enables the following intuitive, trivial adoptions of (termination criteria variant I) derives a Boolean formula encoding all MHS candidates with \( 1 \leq |MHS| \leq b \).

**Lemma 3** Given a bound \( b \), \( H(C, 0, b) \) using the following adapted rules (termination criteria variant I) derives a Boolean formula encoding all MHSs with \( 1 \leq |MHS| \leq b \):

- **RB1:** \( H(False, \ell, b) = True \), \( H(True, \ell, b) = False \);
- **RB2:** \( H(\bar{e}, \ell, b) = True \) if \( \ell \geq b \), else \( e \);
- **RB3:** \( H(e \lor C, \ell, b) = False \) if \( \ell \geq b \), else \( e \lor H(C, \ell, b) \);
- **RB4:** \( H(e \lor C, \ell, b) = False \) if \( \ell \geq b \), else \( e \land H(C, \ell + 1, b) \);
- **RB5:** \( H(C, \ell, b) = False \) if \( \ell \geq b \), else \( e \lor H(C, \ell + 1, b) \lor H(C_2, \ell, b) \).

Proof. The correctness follows from that of the original algorithm and the following considerations regarding a branch’s potential for adding elements to the intermediate solution possibly resulting in violations of the new postcondition regarding bound \( b \). Obviously, \( e \) corresponds exactly to the size of a branch’s intermediate solution, and returning False in some rule would remove the current branch from consideration. R1 cannot add to the solution, so it is left unchanged for RB1. For R2 that adds an element, any \( \ell \geq b \) would result in a breach of bound \( b \), so that we return False then in RB2. Also R3 would add at least one element in any case (the recursion would be in RB3 or RB2), so that we return False in RB3 for \( \ell \geq b \). The same is obvious for RB4 and RB5 (where for R5 the right branch would add at least one element in the recursion to RB2-RB3). Thus only those branches leading to solutions violating the post-condition are removed from the original computation.

Intuitively, if the length of an intermediate solution has reached the bound \( b \), only R1 has to be considered for refinement. Any other rule would result in adding at least one further element to the current branch. Therefore, then only \( e_1 \land e_2 \land \cdots \land e_n \land H(False, b, b) \) may lead to an MHS of length \( b \).

Taking a closer look at situations \( H(C, b - 1, b) \) (i.e. we can add only one further element), unveils the potential for another optimization, as then only those elements present in all \( CS \) are of interest.

**Lemma 4** Let RB1 to RB4 be as RB1 to RB4, and RBS' as follows (termination criteria variant I). Then \( H(C, 0, b) \) derives a Boolean formula encoding all MHSs with \( 1 \leq |MHS| \leq b \).

\[
RBS': \quad H(C, \ell, b) = \begin{cases} 
    False & \text{if } \ell \geq b \lor \exists \ \
    \lor \forall e \lor e_j \\
    \lor \exists \ \
    \lor \forall e \lor e_j \\
    \lor (C_1, \ell + 1, b) \lor H(C_2, \ell, b)
    \end{cases} \\
    \text{with } I = \bigcap_{CS \in C} C_1, \text{ and } C_1, C_2 \text{ as defined previously.}
\]

Proof. This lemma’s correctness follows directly from that for variant I and the fact that for \( \ell = b - 1 \) only and exactly those elements in the intersection of all \( CS \) in \( C \) can complete the current decisions to encode further MHS candidates.

**Corollary 1** No combination of the proposed modifications in the form of Lemmas 1 to 4 and adopting H2 does affect the correctness of the algorithm.

Proof. The corollary’s correctness directly follows from the individual proofs and the absence of preconditions.

4 EVALUATION

In our experiments, we evaluated both run-time and memory consumption of the Boolean approach in various configurations implemented in Python (cf. [8] for an evaluation of some MHS algorithm implementations in Java and Python), comparing it against Reiter’s HS-DAG [9, 5] which we also implemented in Python. We investigated whether any advantages would manifest only under certain conditions, and report a conclusive selection of our results.

Regarding an implementation, there is the obvious option of a recursive approach, but one might also consider an iterative implementation that maintains a list of intermediate work-packages (containing tangles of intermediate Cs (SCSs) and established solution parts) to be iteratively resolved. Some advantage would be the option to be able to (stop and) continue the computation (i.e. with an increased cardinality limit). For our evaluation we implemented both options.

Test Scenario TS1: Artificial random CSs s ∈ SCS. In this scenario, an SCS contains elements drawn randomly from a set of components, i.e. every component is included in a CS, with a probability of 0.5. For this random setting we evaluated the bounded and unbounded case, with a well-sized sample set in order to avoid any bias from a specific random pattern.

Test Scenarios TS2 to TS4: Real-world scenarios based on ISCAS benchmarks. To evaluate whether our findings from TS1 would transfer to real applications, for TS2 to TS4 we constructed SCSs as they would occur in logic circuit diagnosis for the ISCAS’85 benchmark suite [6]. This suite provides ten circuits such as interrupt controllers, modules for single-error-correction (SEC), double-error-detection (DED) and arithmetic logic units (ALU) with 160 to 3512 gates. Our test scenarios were constructed from the circuits c499.isc (32bit SEC, TS2), c880.isc (8bit ALU, TS3), and c1355.isc (32bit SEC, TS4) having 41/60/41 inputs, 32/100/32 outputs, and 202/383/546 gates respectively. Equipping every gate \( g \) with a behavioral assumption \( AB(g) \) that encodes whether it operates correctly or not, we defined a SAT Problem \( P \) of the form

\[
P = \bigwedge_{g_i \in G} (\neg AB(g_i) \Rightarrow output_i := f_{g_i}(in_{g_i}^1, in_{g_i}^2, \ldots))
\]

where \( G \) is the set of gates, and \( output_i, \) \( in_{g_i}^1, \) and \( f_{g_i} \) are a gate \( g_i \)'s output/input signals and its Boolean function. The \( CS \) s are those unsatisfiable cores of these assumptions that violate the original, contradicting input-output observations. For our scenarios, we purposefully injected a single fault by altering the logic function of a gate and computed SCS as the set of \( CS \) s derived during a cardinality-restricted \( |MHS| \leq 3 \) on-the-fly diagnosis run using HS-DAG and Yices (http://yices.csl.sri.com) as theorem prover.
4.1 Experimental results

Our experiments were executed on a 2011 generation MacBook Pro (Intel Core i5 CPU, 2.3 GHz, 4GiB RAM, SSD, Mac OS X 10.6) using CPython 2.7.1. With swapping and the GUI disabled, each sample faced resource limits of 300 seconds and 2GiB RAM. Regarding memory usage, we polled the process’ resident set size (RSS) from the operating system in a separate process, and report its maximum. We plot run-times and memory usage on logarithmic scales, with approximately 120 points equally distributed on the x-axis (e.g. for Figure 1) we have \( |SCS| \approx 10^{6i/120} \) for \( 0 \leq i \leq 120 \).

Figure 1 shows run-time and maximum RSS averaged over 20 samples for TS1 in the unbounded case with \( |COMP| = 20 \) and a varying \( |SCS| \). V1 denotes using H1, while V2 amounts to using H2, and V3 implements Lemma 2. R4 vs. R4’ should be self-explanatory, while Rec indicates a recursive implementation and It an iterative one. The first thing to observe is that all Boolean variants are about one order of magnitude faster than HS-DAG in the range \([10,300]\) and also consume significantly less memory. While using R4’ ensures this also for larger SCS, the performance advantage diminishes otherwise, so that HS-DAG can outperform the Boolean approach when an SCS gets larger than approx. \( 6 \cdot 10^3 \). While variants V2 and V3 suffer minor drawbacks for small \( |SCS| \), they slightly gain in performance for higher amounts of CSs. All variants using R4’ feature very similar memory characteristics. Using a recursive or iterative implementation does not result in huge differences, with the recursive ones being a touch faster and the iterative ones a bit more memory-effective. While we were motivated mostly by issues with bounded computations, these results suggest that there is little to no computation overhead in the unbounded case, and specifically R4’ can also help in an unbounded search. Please note that the missing plot segments for HS-DAG stem from (run-time) limit violations.

Figure 2 shows our results for TS1 and \( |MHS| = 1 \). Those variants implementing variant II of the termination criteria are indicated by the suffix “Stop”, all others use variant I. The graphs at the top plot the recursive implementations' performance and illustrate that, as expected, R4’ has no influence on time or memory usage at all (due to the very small probability of SCS containing a CS of size one). Compared to V1 using termination criteria I, variants V2 and V3 already allow a significant boost, surpassed however by that offered by termination criteria II. For the latter the strategy also does not play a significant role, which is intuitive given the small amount of recursions allowed by those criteria. Obviously, as in this case Rule 5 is never really executed (in fact, only the intersection of all conflict sets is computed), all heuristics share the same time- and memory characteristics. Since our recursive and iterative implementations again perform similarly, we will not plot the iterative ones for our further figures due to space restrictions. Overall, while HS-DAG’s performance is still superior for \( |SCS| \) larger than about \( 10^4 \), the original threshold was below ten. Furthermore, we enhanced the performance of the Boolean approach by up to two orders of magnitude.

A last evaluation of TS1 is shown in Figure 3, where we restricted \( |MHS| \) to 3. Like for simple samples in the unbounded case, we encountered a negative impact of R4’ (for the majority of the smaller samples) when using the original heuristic V1. For the other variants the impact of R4’ diminishes, specifically when using termination criteria II. Again V3 outperforms V2, that in turn outperforms V1. Also for these tests, we could significantly raise the border where HS-DAG would take the lead from about 20 to approx. 200.

Finally, Figure 4 shows the performance of several implementations for 300 real-world samples (100 samples for each of TS2 to TS4) with cardinality limits \{1, 2, 3\}. Due to random fault injection, samples with similar \( |SCS| \) may still have incommensurable structural complexity, leading to a large variance in both run-time and memory. In order to observe otherwise obscured trends, we applied a moving average filter that considers for any \( |SCS| \) \( x_0 \) on the x-axis all samples within the range \( x_0/\sqrt{2}, x_0\sqrt{2} \). Considering the graphs, we find that our conclusions from the artificial scenario would also transfer to these SCSs as extracted from a real-world application. While for the original heuristic (V1) R4’ results in a run-time penalty, this drawback vanishes for termination criteria II. The variant ranking V3–V2–V1 is confirmed, as, e.g., evident from the graph for \( |MHS| \leq 2 \). However, the most notable result follows from the comparison with HS-DAG. While for \( |MHS| = 1 \) and \( |SCS| \geq 5 \), HS-DAG is faster than the original Boolean variant, our termination criteria II alone (then the variant has almost no impact) leads to an average advantage of one order of magnitude compared to HS-DAG. When raising the maximum cardinality to \( |MHS| = 2 \) and 3, more and more Boolean variants catch up, until for max. \( |MHS| = 3 \) only one of them (V1-R4’) is slower than HS-DAG. Also the ranking V3-V2-V1 becomes more apparent then. Summing up, with our optimizations the Boolean approach is not only a performant contender in the unbounded case, but also for a bounded MHS search.

![Figure 1](image1.png)  
Figure 1. Run-times for TS1, \( |COMP| = 20 \), and an unbounded MHS search.
5 CONCLUSIONS

In this paper we showed how to improve the Boolean approach’s performance by up to 2 orders of magnitude for artificial and real-world scenarios for which it was known to perform badly before. That is, we showed a new heuristic, a revised split strategy, as well as tight termination rules that tackle earlier disadvantages when conquering the search space for cardinality-bounded problems. Our experiments suggest that our optimizations would not significantly hinder performance in the unbounded case, and could sometimes also enhance it. We also proposed a slightly improved Rule 4 that avoids some duplicates and non-minimal solutions to be pruned anyway. Thus, the Boolean algorithm now is also an attractive alternative for the bounded case. In future work, we will try to tackle the Boolean algorithm’s main drawback that it requires SCS to be known in advance, providing an interesting option for our main research regarding enabling model-based diagnosis of formal temporal specifications.

ACKNOWLEDGEMENTS

This work has been funded by the Austrian Science Fund (FWF) under grant P22959-N23. We would like to thank the reviewers for their comments and Franz Wotawa for fruitful discussions.

REFERENCES

Figure 3. Results for TS1, $|COMP| = 20$, max. $|MHS| = 3$.

Figure 4. Results for TS2 to TS4 with varying max. $|MHS|$. 