On the Practicality of Abductive Validation

Tim Menzies [1]

Abstract. An abductive framework is described for validating theories using a library of known or desired behaviour. Abduction is known to be NP-hard which suggests that this framework is impractical for anything other than small theories. The computational limits of the framework is therefore explored. We find that abductive validation is a practical tool for the KBS we see in contemporary practice.

KEYWORDS: Abduction, validation, computational complexity, expert systems, knowledge acquisition.

1 INTRODUCTION

The connection between abduction and other KBS inference tasks (e.g. model-based diagnosis) is well-documented [2, 7]. It would be convenient if we could execute and test our KBS in the same abductive framework. This would remove the need for complicated translations between the executable form of a KBS and its associated test engine.

Here we explore KBS validation using HT4, an abductive inference engine. HT4 assumes that the definitive test for a model is that it can reproduce (or *cover*) known behaviour of the entity being modeled. Theory $\mathcal{T}_1$ is a better theory than theory $\mathcal{T}_2$ if $\mathcal{T}_1^{cover} \gg \mathcal{T}_2^{cover}$. HT4 is a generalisation and optimisation of QMOD, a validation tool for neuroendocrinological theories [5].

One drawback with abduction is that it is slow. Selman & Levesque show that even when only one abductive explanation is required and the theory is restricted to be acyclic, then abduction is NP-hard [9]. Bylander *et. al.* make a similar pessimistic conclusion [1]. Computationally tractable abductive inference algorithms (e.g. [1, 4]) typically make restrictive assumptions about the nature of the theory or the available data. Such techniques are not applicable to arbitrary theories. Therefore, it is reasonable to question the practicality of HT4 for medium to large theories.

This paper is structured as follows. Section 2 introduces abductive validation and Section 3 explores its complexity. Section 4 review studies which experimentally demonstrate the practicality of HT4 for the KBS we see in contemporary practice.

2 ABDUCTIVE VALIDATION

Abduction is the search for assumptions $\mathcal{A}$ which, when combined with some theory $\mathcal{T}$ achieves some set of goals $\mathcal{OUT}$ without causing some contradiction [4]. That is:

- EQ_1: $\mathcal{T} \cup \mathcal{A} \vdash \mathcal{OUT}$;
- EQ_2: $\mathcal{T} \cup \mathcal{A} \not\vdash \bot$.

HT4 caches the proof trees used to satisfy EQ_1 and EQ_2. These are then sorted into *worlds* $\mathcal{W}$: maximal consistent subsets (maximal with respect to size). Each world condones a set of inferences. A world's *cover* is the size of the overlap between $\mathcal{OUT}$ and that world.

For example, given a set of goal $\mathcal{OUT}$puts and known $\mathcal{IN}$puts, then HT4 can use (e.g.) the qualitative theory of Figure 1 to build a set of proof trees $\mathcal{P}$ connecting $\mathcal{IN}$puts to $\mathcal{OUT}$puts. In Figure 1, (i) $\texttt{x} \overset{++}{\rightarrow} \texttt{y}$ denotes that `y` being `up` or `down` could be explained by `x` being `up` or `down` respectively while (ii) $\texttt{x} \overset{--}{\rightarrow} \texttt{y}$ denotes that `y` being `up` or `down` could be explained by `x` being `down` or `up` respectively. If we assume that (i) the conjunction of an `up` and a `down` can explain a `steady` and that (ii) no change can be explained in terms of a `steady` (i.e. a `steady` vertex has no children), then we can partially evaluate Figure 1 into the and-or graph of literals shown in Figure 2. This graph contains one vertex for each possible state of the nodes of Figure 1 as well as *and* vertices which models combinations of influences (for example, `gDown` and `bDown` can lead to `fSteady`). The dotted lines in Figure 1 denote edges around *and* vertices.

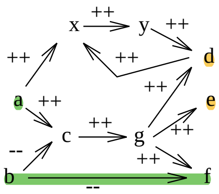

Figure 1. A qualitative theory.

For example, in the case where $\mathcal{OUT}$ = {`dUp,eUp,fDown`} and $\mathcal{IN}$={`aUp,bUp`}, then $\mathcal{P}_1$=`aUp` $\rightarrow$ `xUp` $\rightarrow$ `yUp` $\rightarrow$ `dUp`, $\mathcal{P}_2$= `aUp` $\rightarrow$ `cUp` $\rightarrow$ `gUp` $\rightarrow$ `dUp`, $\mathcal{P}_3$= `aUp` $\rightarrow$ `cUp` $\rightarrow$ `gUp` $\rightarrow$ `eUp`, $\mathcal{P}_4$= `bUp` $\rightarrow$ `cDown` $\rightarrow$ `gDown` $\rightarrow$ `fDown`, $\mathcal{P}_5$= `bUp` $\rightarrow$ `fDown`.

Some of these proofs make assumptions; i.e. use a literal that is not one of the known $\mathcal{FACTS}$ (typically, $\mathcal{FACTS} = \mathcal{IN} \cup \mathcal{OUT}$). Note that some of the assumptions will contradict other assumptions and will be *controversial* (denoted $\mathcal{A}_C$). For example, assuming `cDown` and `cUp` at the same time is contradictory. DeKleer's key insight [3] was that, in terms of uniquely defining an assumption space, the key controversial assumptions are those controversial assumptions that are not dependent on other controversial assumptions. We denote these *base controversial* assumptions $\mathcal{A}_B$. In our example, $\mathcal{A}_C$={`cUp,cDown,gUp,gDown`} and $\mathcal{A}_B$ = {`cUp, cDown`} (since Figure 1 tells us that `g` is fully determined by `c`). If we assume `cUp`, then we can believe in the *world*

[1] Department of Artificial Intelligence, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia `timm@insect.sd.monash.edu.au` `http://www.sd.monash.edu.au/~timm`

ECAI 96. *12th European Conference on Artificial Intelligence*
Edited by W. Wahlster
Published in 1996 by John Wiley & Sons, Ltd.

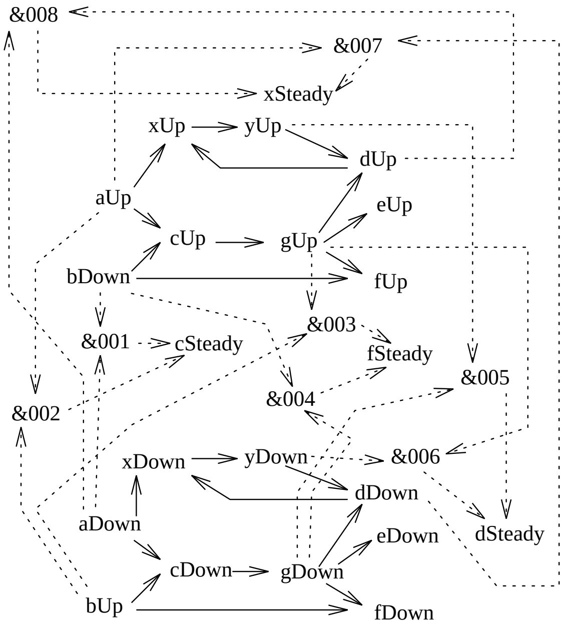

Figure 2. The search space tacit in Figure 1

$\mathcal{W}_1$ containing the proofs $\mathcal{P}_1$ $\mathcal{P}_2$ $\mathcal{P}_3$ $\mathcal{P}_5$ since those proofs do not assume `cUp`. If we assume `cDown`, then we can believe in the world $\mathcal{W}_2$ containing the proofs $\mathcal{P}_1$ $\mathcal{P}_4$ $\mathcal{P}_5$ since these proofs do not assume `cDown`. These worlds are shown in Figure 3.

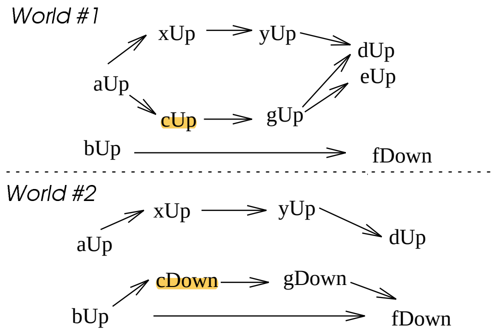

Figure 3. Two worlds from Figure 2

The overlap of $\mathcal{W}_1$ and $\mathcal{OUT}$ is {`dUp,eUp,fDown`} and the overlap $\mathcal{W}_2$ and $\mathcal{OUT}$ is {`dUp,fDown`}; i.e. $\mathcal{W}_1^{cover}$ = 3 = 100% and $\mathcal{W}_2^{cover}$ = 2 = 67%. The maximum cover is 100%; i.e. (i) their exist a set of assumptions ({`cUp`}) which let us explain all of $\mathcal{OUT}$ and (ii) this theory passes abductive validation.

3 COMPLEXITY OF HT4

The core problem in HT4 is finding $\mathcal{A}_B$. In the *forward sweep*, HT4 finds $\mathcal{A}_C$ as a side-effect of computing the transitive closure of $\mathcal{IN}$. In the *backwards sweep*, HT4 constrains proof generation to the transitive closure of $\mathcal{IN}$. As a proof is grown from a member of $\mathcal{OUT}$ back to $\mathcal{IN}$, five invariants are maintained. (i) Proofs maintain a *forbids* set; i.e. a set of literals that are incompatible with the literals used in the proof. For example, the literals used in $\mathcal{P}_1$ forbid the literals {`aDown, aSteady, xDown, xSteady, yDown, ySteady, dDown, dSteady` }. (ii) A proof must not contain loops or items that contradict other items in the proof (i.e. a proof's members must not intersect with its *forbids* set). (iii) If a proof crosses an *and* node, then all the parents of that node must be found in the proof. (iv) A literal in a proof must not contradict the known $\mathcal{FACTS}$. (v) The upper-most $\mathcal{A}_C$ found along the way is recorded as that proof's *guess*. The union of all the guesses of all the proofs is $\mathcal{A}_B$.

Once $\mathcal{A}_B$ is known then the proofs can be sorted into worlds in the *worlds sweep*. HT4 extracts all the objects $\mathcal{O}$ referenced in $\mathcal{A}_B$. A world-defining environment $\mathcal{ENV}_i$ is created for each combination of objects and their values. In our example, $\mathcal{ENV}_1$ = {`cUp`} and $\mathcal{ENV}_2$ = {`cDown`}. The worlds sweep is simply two nested loops over each $\mathcal{ENV}_i$ and each $\mathcal{P}_j$. A proof $\mathcal{P}_j$ belongs in world $\mathcal{W}_i$ if its *forbids* set does not intersect the assumptions $\mathcal{ENV}_i$ that define that world. For more details on the internals of HT4, see [7].

HT4's runtimes are clearly exponential on theory size. In a theory comprising a directed and-or graph connecting literals $\mathcal{V}$ with $\mathcal{E}$ edges and average fan-in $\mathcal{F}=\frac{|\mathcal{E}|}{|\mathcal{V}|}$, the worst-case complexity of the forwards sweep is acceptable at $O(|\mathcal{V}|^3)$. However, if the average size of a proof is X, then worse case backwards sweep is $O(X^{\mathcal{F}})$. Further, the worlds sweep is proportional to the number of proofs and the number of world-defining assumptions; i.e. (i.e. $O(|\mathcal{P}| * |\mathcal{ENV}|)$ = $O(|\mathcal{X}^{\mathcal{F}}| * |\mathcal{ENV}|)$.

4 EXPERIMENTS

4.1 The Smythe '87 Study

The Smythe '87 [10] theory shown in Figure 4 proposes connections between serum adrenocorticotropin (`acth`), serum corticosterone (`cortico`), and neuro-noradrenergic activity (`nna`). `Nna` was measured as the ratio of noradrenaline to its post-cursor, 3,4-dihydroxphenyl-ethethyleneglycol.

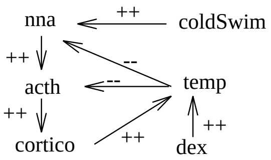

Figure 4. The Smythe '87 theory.

The Smythe '87 paper contains values comparing these values in various experiments: (i) *Control* i.e. no treatments; (ii) `dex` i.e. an injection of dexamethasone at 100 $\frac{mg}{kg}$; (iii) `coldSwim` i.e. a two minute swim in a bath of ice cold water; and (iv) `coldSwim, dex` i.e. both a `coldSwim` and an injection of `dex`. A sample of experimental results from Smythe '87 is shown in Figure 5. The experiments are shown in the columns and the measures taken in the different experiments are shown in the rows.

In the comparison `coldSwim` to `dex, coldSwim`, $\mathcal{OUT}$={`acthDown, corticoDown, nnaUp`}. If we model `dex`

Figure 5. Some results from Smythe '87.

Measured	`coldSwim`	`dex`	`dex,`
			`coldSwim`

`nna`	0.210	0.105	0.246
`cortico`	1231	11.3	32.8
`acth`	240	0	0

and `coldSwim` as booleans with values $< 0, 1 >$ then for this comparison, $\mathcal{IN}$={`dexUp`}. In this comparison `nnaUp` can't be explained since their exists no proof from `nnaUp` to $\mathcal{IN}$ which does not violate the proof invariants. Another error can be found in the comparison `dex` to `dex,coldSwim`. In this comparison $\mathcal{IN}$={`coldSwimUp`} and $\mathcal{OUT}$={`acthSteady, corticoUp, nnaUp`} and only `nnaUp` can be explained.

Note that the faults of this theory were found by a detailed examination of the data published to support it. Further, the errors were not known by the author of the theory, till we pointed it out to him. Lastly, these errors escaped international peer review.

4.2 Smythe '89

.

Smythe '87 is a small theory. The Smythe '89 study explores how well abductive validation scales up to medium-sized models.

Smythe '89 [11] is a theory of human glucose regulation. It contains 27 possible inputs and 53 measurable entities which partially evaluated into an and-or graph with $|\mathcal{V}|$ = 554 and $\frac{|\mathcal{E}|}{|\mathcal{V}|} = 2.25$. Smythe '89 is a review paper that summaries a range of papers from the field of neuroendocrinology. Those papers offer 870 experimental comparisons with between 1 to 4 inputs and 1 to 10 outputs.

Smythe '89 was originally studied by QMOD. That study found that 32% of the observations were inexplicable. QMOD could not explain studies or handle multiple causes. These restrictions implied that it could only handle of 24 possible comparisons. Even with these restrictions, QMOD found several errors in Smythe '89 that were novel and exciting to Smythe himself [5]. Like the Smythe '87 study, these errors had not been detected previously by international peer review.

When HT4 ran over the full 870 comparisons, it found more errors than QMOD. Only 150 of the comparisons could explain 100% of their $\mathcal{OUT}$puts. On average, 45% of the $\mathcal{OUT}$s in those comparisons were inexplicable. The level of critique offered by QMOD and HT4 is surprisingly high. This is both a disturbing and exciting finding. It is disturbing in the sense that if the very first large-scale medical theory analysed by abductive validation contains significant numbers of errors, then it raises doubts as to the accuracy of theories in general. This result is exciting in the sense that the level of critique is so high. Abductive validation promises to be a powerful tool for assessing vague theories.

4.3 The Mutation Study

The previous studies are interesting, but a sample size of 2 is inadequate to make general claims. In the mutation study, 257 theories were artificially generated by adding random vertices and edges to the and-or graph from Smythe '89. These were run using 5970 experimental comparisons.

For the *changing I/O* study, 1391 runs were made across the Smythe '89 model with an increasing number of $\mathcal{IN}$puts and $\mathcal{OUT}$puts. Surprisingly, HT4 runtimes were noted to decrease as the size of $\mathcal{IN}$ and $\mathcal{OUT}$ was increased. This seemed a surprising result till we realised that every $\mathcal{FACT}$ culls the search space. For example, if know that `day=tuesday`, then that rules out `day=monday, day=wednesday,`... etc.

In the *changing N* mutation study, the fanout was kept constant and the number of vertices were increased. Figure 6 shows the average runtime for executing HT4 over 94 and-or graphs and 1991 $< \mathcal{IN}, \mathcal{OUT} >$ pairs [7]. For that study, a "give up" time of 840 seconds was built into HT4. HT4 did not terminate for $|\mathcal{V}| \geq 850$ in under that "give up" time (shown in Figure 6 as a vertical line). We conclude from Figure 6 that the "knee" in the exponential runtime curve kicks-in at around 800 literals. These result came from a Smalltalk V implementation on a Macintosh Powerbook 170. A port to "C" on a Sparc Station is underway.

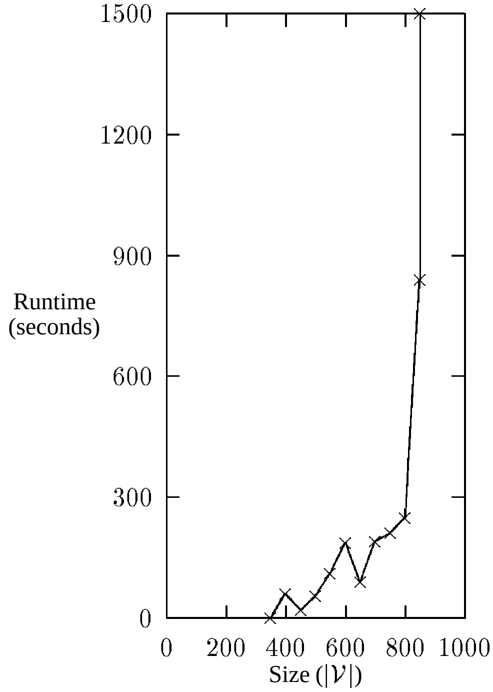

Figure 6. Average runtimes.

In practice, how restrictive is a limit of 850 vertices? Details of the nature of real-world expert systems are hard to find in the literature. The only reliable data we could find is shown in Figure 7 which shows the size of the dependency graph between literals in fielded propositional expert systems [8]. Figure 7 suggests that a practical inference engine must work at least for the range $55 \geq |\mathcal{V}| \geq 510$ and $2 \geq \frac{|\mathcal{E}|}{|\mathcal{V}|} \geq 7$.

Note that the Figure 6 results were obtained from a less-than-optimum platform: Smalltalk running on a Macintosh. However, the current results on a relatively slow platform show that HT4 is practical for the theory sizes we see in practice.

The *changing fanout* mutation study examined the practicality of HT4 for models of varying fanout. In that study, the Smythe '89

Figure 7. Figures from fielded expert systems.

Application	$\lvert\mathcal{V}\rvert$	$\frac{\lvert\mathcal{E}\rvert}{\lvert\mathcal{V}\rvert}$
displan	55	2
mmu	65	7
tape	80	4
neuron	155	4
DMS-1	510	6

theory size was kept constant, but edges were added at random to produce new graphs of larger fanouts. Six models were used of sizes $|\mathcal{V}| = \{449, 480, 487, 494, 511, 535\}$. Figure 8 shows the results. At low fanouts, many behaviours were inexplicable. However, after a fanout of 4.4, most behaviours were explicable. Further, after a fanout of 6.8, nearly all the behaviours were explicable [7].

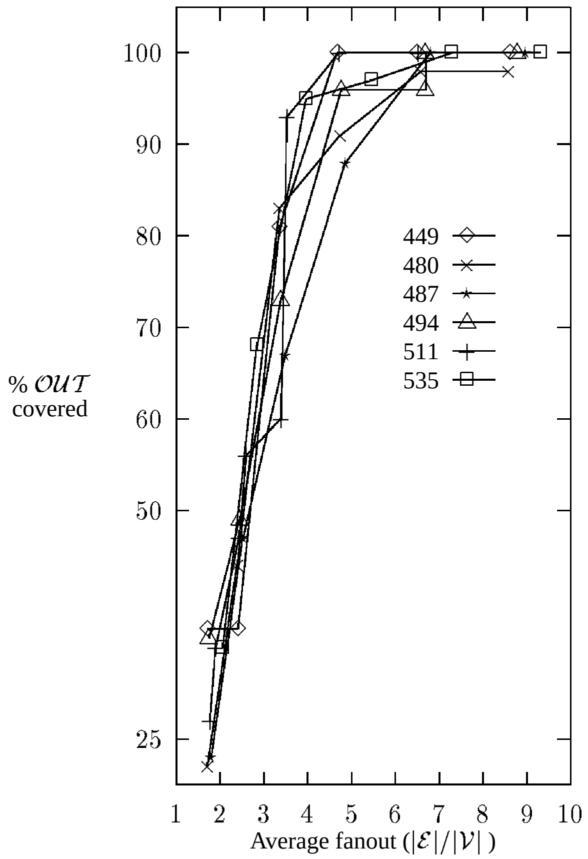

Figure 8. Explicable outputs.

As a result of the mutation study we conclude that HT4 is practical for the size and fanout of expert systems seen in current practice and for nearly the range of fanouts seen in fielded expert systems. However, after a certain level of inter-connectivity, a theory is able to reproduce any input/output pairs. An inference procedure that condones any behaviour at all from a theory is not a useful inference procedure. After the point where % $\mathcal{OUT}$ *covered* approaches 100%, the HT4 becomes a useless validation tool.

5 DISCUSSION

In the literature, there at least two types of KBS testing: verification and validation. Verification tools search for syntactic anomalies within a knowledge base such as tautologies, redundancies, and circularities in the dependency graph of literals in a knowledge base [8]. Verification is not a definitive test for a KBS. Preece reports example where working expert systems contained syntactic anomalies, yet still performed adequately [8]. Validation tools assess a knowledge via some external semantic criteria; e.g. testing that a knowledge base model of X can reproduce known behaviour of X. If such a test suite of behaviour is missing, then non-monotonic reasoning techniques can be used to explore the dependency graph between KB literals to find sets of input literals which will exercise the entire knowledge [6, 12].

In this paper, we have explored a non-monotonic abductive validation variant which assumes the presence of a library of known behaviour. HT4-style abductive validation executes over a finite and-or graph of literals. Many representations can be mapped into this form. Our examples here are from qualitative theories but the technique could be applied to propositional rule-bases or any first-order theory that can be unfolded in a finite number of steps to a ground theory. Abductive validation handles certain hard and interesting cases; e.g. the processing of indeterminate models where inference implies handling mutually exclusive assumptions in different worlds. We have shown examples where this framework has faulted theories published in the international peer-reviewed literature. Interesting, we have found these faults using the data published to support those theories.

We have explored the computational limits of this approach and concluded that abductive validation can handle at least the KBS systems seen in current practice (as defined by the Figure 7 survey).

ACKNOWLEDGEMENTS

The author wished to thank A/Prof. Paul Compton (University of New South Wales) for his support of this project.

REFERENCES

[1] T. Bylander, D. Allemang, M.C. M.C. Tanner, and J.R. Josephson. The Computational Complexity of Abduction. *Artificial Intelligence*, 49:25–60, 1991.

[2] L. Console and P. Torasso. A Spectrum of Definitions of Model-Based Diagnosis. *Computational Intelligence*, 7:133–141, 3 1991.

[3] J. DeKleer. An Assumption-Based TMS. *Artificial Intelligence*, 28:163–196, 1986.

[4] K. Eshghi. A Tractable Class of Abductive Problems. In *IJCAI '93*, volume 1, pages 3–8, 1993.

[5] B. Feldman, P. Compton, and G. Smythe. Hypothesis Testing: an Appropriate Task for Knowledge-Based Systems. In *4th AAAI-Sponsored Knowledge Acquisition for Knowledge-based Systems Workshop Banff, Canada*, 1989.

[6] A. Ginsberg. Theory Reduction, Theory Revision, and Retranslation. In *AAAI '90*, pages 777–782, 1990.

[7] T.J. Menzies. Applications of Abduction: Knowledge Level Modeling. Technical Report TR95-23, Department of Software Development, Monash University, 1995. To appear in the *International Journal of Human Computer Studies*, August, 1996.

[8] A.D. Preece and R. Shinghal. Verifying Knowledge Bases by Anomaly Detection: An Experience Report. In *ECAI '92*, 1992.

[9] B. Selman and H.J. Levesque. Abductive and Default Reasoning: a Computational Core. In *AAAI '90*, pages 343–348, 1990.

[10] G.A. Smythe. Hypothalamic noradrenergic activation of stress-induced adrenocorticotropin (ACTH) release: Effects of acute and chronic dexamethasone pre-treatment in the rat. *Exp. Clin. Endocrinol. (Life Sci. Adv.)*, pages 141–144, 6 1987.

[11] G.A. Smythe. Brain-hypothalmus, Pituitary and the Endocrine Pancreas. *The Endocrine Pancreas*, 1989.

[12] N. Zlatereva. Truth Mainteance Systems and Their Application for Verifying Expert System Knowledge Bases. *Artificial Intelligence Review*, 6, 1992.