Random vs. Best-First: Impact of Sampling Strategies on Decision Making in Model-Based Diagnosis

Patrick Rodler, University of Klagenfurt, Austria

patrick.rodler@aau.at

1 Model-Based Diagnosis (MBD)

An important task in MBD is the efficient localization of faults

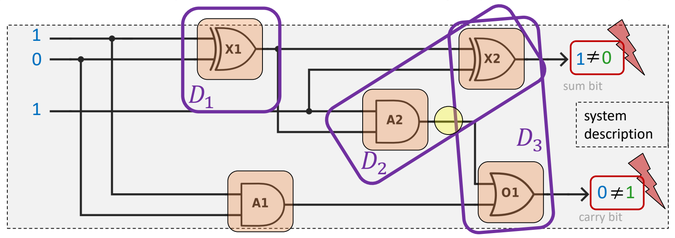
Fault Localization:
Given: system (e.g., SW, HW, KB, physical device, CSP, ontology, etc.)
- consisting of a set of components (e.g., lines of code, gates, logical sentences)
- which does not behave as expected

Find: the faulty components that cause the misbehavior

Example: Full-Adder below does not add properly

Find diagnosis (⊆-minimal set of components that, when assumed faulty, explains misbehavior) !

Components
Observations
Predictions (by theorem prover, based on system description + assumption that all components are normal)
Discrepancy (Predictions vs. Observations)



Multiple diagnoses! Which one is the correct fault?
1. Use diagnosis probabilities (determined based on the likeliness of component failure)
2. Apply **Sequential Diagnosis** to localize the correct fault with certainty

2 Sequential Diagnosis

Example (cont'd): Which diagnosis among $\boldsymbol{D} = \{D_1, D_2, D_3\}$ is the actual fault?

Collect further information to rule out spurious diagnoses → make measurements

E.g.: Measurement point (MP) $out(A_2)$ is informative wrt. $\boldsymbol{D}$
→ if outcome is 0, then D_3 is no longer a diagnosis
→ if outcome is 1, then D_1, D_2 are no longer diagnoses

eliminates at least one diagnosis in $\boldsymbol{D}$, regardless of the measurement outcome

General Process: heuristics are used since optimal MP selection is NP-hard
- Conduct measurements until a single (highly probable) diagnosis remains
- Always select best informative MP → "best" defined based on a MP selection heuristic
- Basis for MP selection = computed set of diagnoses $\boldsymbol{D}$ + diagnosis probabilities
- Diagnoses + probabilities allow to estimate
  - probability of different measurement outcomes, and
  - (rate of) eliminated diagnoses for different measurement outcomes

common heuristics evaluate MPs based on exactly these two factors

3 Motivation & Contribution

Assume an election poll:
- Ask only university professors for whom they will vote
- Will the result of the poll be representative of the entire population?

Similar thing is often done in MBD
- Task: find actual diagnosis among a (large) set of diagnoses
- Computing all diagnoses intractable → compute only a sample of diagnoses
- Use sample to make estimations that guide diagnostic actions (meaurements)
- Draw best-first samples (e.g. most probable diagnoses)

But:
Statistical Law:

"A randomly chosen unbiased sample from a population allows (on average) better conclusions and estimations about the whole population than any other sample."

Questions of Interest:
- Does this apply to MBD as well?
- Or are best-first samples really more informative than random ones in MBD?
- Can we do better by using randomized algorithms to generate diagnoses?

Contribution:
Comprehensive empirical evaluations to bring light to these questions

References

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. Artif. Intell. 32(1), 57-95 1987.
[Rodler, 2015] Patrick Rodler. Interactive Debugging of Knowledge Bases. PhD Thesis, Univ. Klagenfurt. 2015.
[Schekotihin et al, 2014] Konstantin Schekotihin, Gerhard Friedrich, Patrick Rodler, Philipp Fleiss. Sequential diagnosis of high cardinality faults in knowledge-bases by direct diagnosis generation. In: ECAI. 2014.

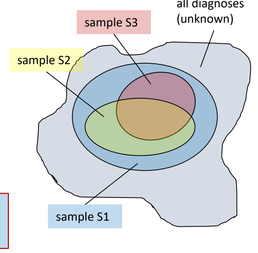
4 Example (Impact of Sample in MBD)

Diagnosis problem:
4 diagnoses: $D_1, \dots, D_4$
probabilities: $\langle p(D_1), \dots, p(D_4) \rangle = \langle .37, .175, .175, .28 \rangle$
Consider MPs m_1, m_2 → two possible outcomes (T/F) each
Given a sample of diagnoses → assess quality of each MP based on its properties wrt.
- probability p of T/F outcomes
- diagnosis elimination rate e for T/F outcomes

Consider 3 samples $S1 = \{D_1, D_2, D_3, D_4\}, S2 = \{D_1, D_2, D_3\}, S3 = \{D_2, D_3, D_4\}$

1. Different samples can yield significantly different estimations

$p_{S1}(m_1 = T) = .55$ $p_{S1}(m_1 = F) = .45$
$p_{S2}(m_1 = T) = .76$ $p_{S2}(m_1 = F) = .24$
$e_{S1}(m_1 = T) = .5$ $e_{S1}(m_1 = F) = .5$
$e_{S2}(m_1 = T) = .33$ $e_{S2}(m_1 = F) = .67$



→ similar observations for other MP selection heuristics!

2. Different samples can lead to different diagnostic decisions

$p_{S1}(m_1 = T) = .55$ $p_{S1}(m_1 = F) = .45$
$p_{S1}(m_2 = T) = .72$ $p_{S1}(m_2 = F) = .28$
m_1 better wrt. information gain heuristic

$p_{S3}(m_1 = T) = .28$ $p_{S3}(m_1 = F) = .72$
$p_{S3}(m_2 = T) = .55$ $p_{S3}(m_2 = F) = .45$
m_2 better wrt. information gain heuristic

5 Evaluation Approach

Dataset: Real-world diagnosis cases (domain: KB/ontology debugging)

specific samples: outcome a-priori known, usually more costly

Sample Types:
- best-first (bf) — most probable diagnoses
- random (rd) — unbiased random selection from all diagnoses
- worst-first (wf) — least probable diagnoses
- approx best-first (abf)
- approx random (ard)
- approx worst-first (awf)

heuristic approximations

unspecific samples: outcome a-priori unknown, usually less costly

# of components | all beyond NP-complete

KB K	$\lvert K \rvert$	expressivity	#D/min/max
University (U)	50	$\mathcal{SOIN}^{(D)}$	90/3/4
IT	140	$\mathcal{SROIQ}$	1045/3/7
UNI	142	$\mathcal{SROIQ}$	1296/5/6
MiniTambis (M)	173	$\mathcal{ALCN}$	48/3/3
Transportation (T)	1300	$\mathcal{ALCH}^{(D)}$	1782/6/9
Economy (E)	1781	$\mathcal{ALCH}^{(D)}$	864/4/8
DBpedia (D)	7228	$\mathcal{ALCHF}^{(D)}$	7/1/1
Cton (C)	33203	$\mathcal{SHF}$	15/1/5

complexity of theorem proving

# of diagnoses
size of smallest diagnosis
size of largest diagnosis

Computation of Samples:

bf	uniform-cost HS-Tree	HS-Tree [Reiter, 1987]; Uniform-Cost HS-Tree [Rodler, 2015]; Inv-HS-Tree [Schekotihin et al, 2014]
rd	compute all diagnoses, sample randomly	
wf	compute all diagnoses, select least probable diagnoses	
abf	Inv-HS-Tree with sorting of components by probability in descending order	
ard	Inv-HS-Tree with random sorting of components	
awf	Inv-HS-Tree with sorting of components by probability in ascending order	

Evaluation Criteria for Sample Types:

Theoretical Representativeness: sample type is the more representative, the better the
- probability estimates for MPs m match real probabilities for m
- elimination rate estimates for MPs m match real elimination rate for m

accuracy

Practical Representativeness: sample type is the more representative, the lower the
- # of measurements
- time

required until the isolation of the actual diagnosis

efficiency

Research Questions:

RQ1: Which type of sample is best in terms of theoretical representativeness?
RQ2: Which type of sample is best in terms of practical representativeness?
RQ3: Are the results wrt. RQ1 and RQ2 consistent over different *(a)* sample sizes, *(b)* MP selection heuristics, and *(c)* diagnosis problem cases?
RQ4: Does larger sample size (more computed diagnoses) imply better representativeness?
RQ5: Does a better theoretical representativeness translate to a better practical representativeness?

Special Focus on:
- Statistical unfoundedness of best-first samples (commonly used in MBD)
- Theoretical attractiveness of random samples (not commonly used in MBD)

6 Two Experiments

EXP1 (theoretical representativeness): 8 x 6 x 5 x 50 = 12.000 MPs; 24.000 (probability + elimination rate) estimates

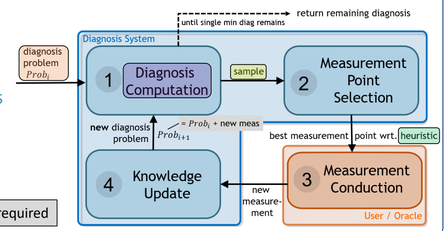
For each of 8 diagnosis cases, 6 sample types, 5 sample sizes in $\{2,6,10,20,50\}$:
→ we computed sample of diagnoses S
→ we randomly selected 50 MPs (if existent) for S
→ we computed probability + elimination rate estimates for each of the 50 MPs
- by means of S → sample estimate
- by means of all diagnoses → "real" value

of interest: comparison of these values

EXP2 (practical representativeness): 8 x 6 x 5 x 4 = 960 factor combinations, 960 x 10 = 9.600 sequential diagnosis sessions

For each of
- 8 diagnosis cases,
- 6 sample types,
- 5 sample sizes in $\{2,6,10,20,50\}$,
- 4 MP selection heuristics:

→ we executed 10 sequential diagnosis sessions until a single diagnosis was isolated
→ we randomly selected the actual diagnosis to be found in each session



of interest: runtimes (sampling, overall) and # measurements required

7 Results

RQ1: Which type of sample is best in terms of theoretical representativeness?

Elimination rate estimations:
rd best (unsurprisingly) --- estimations altogether fairly OK for all sample types --- approximate methods (abf, awf, ard) produced less representative samples than exact ones

Probability estimations:
bf best, rd only 2nd-best --- estimations altogether fairly OK for all sample types --- probability estimations in general less reliable than elimination rate estimations

RQ2: Which type of sample is best in terms of practical representativeness?

Number of measurements:
bf best for heuristics ENT, SPL and small sample size $\{2,6\}$ --- bf worst for heuristic MPS (significant overheads!) --- rd best for heuristic RIO, clearly better than bf for heuristic MPS --- performance of rd depends largely on heuristic --- approximate methods perform quite well (especially ard)

Time:
awf overall best --- rd + wf worst --- in most scenarios: sample type best wrt. time is different from sample type best wrt. # measurements --- in all scenarios: if an exact method (bf, rd, wf) is best wrt. time, then an approximate method (abf, ard, awf) is best wrt. # measurements, and vice versa

Overall time for sequential diagnosis:
bf/abf best for "typical" diagnosis scenarios (i.e., smaller sample size and heuristics ENT, SPL)
rd/ard best for "less typical" diagnosis scenarios (i.e., larger sample size and heuristics RIO, MPS)

RQ3: Results wrt. RQ1/RQ2 stable over different *(a)* sample sizes, *(b)* heuristics, *(c)* problem cases?

RQ1: → overall fairly consistent rankings --- results stable wrt. winning strategy
RQ2: → more fluctuation --- rankings for time tend to be more stable than for # measurements

RQ4: Does larger sample size (more computed diagnoses) imply better representativeness?

Theoretical representativeness: → yes
Practical representativeness: → no (obvious for time, less so for # measurements)

in line with earlier studies

RQ5: Does better theoretical representativeness imply better practical representativeness?

cannot be generally concluded from our results --- possible explanation: approximate nature of heuristic MP evaluation might lower benefit of good estimations

Bottom Line:

given an **efficient** random diagnosis sampling algorithm (open problem!)

Random samples	→ very good estimations	
	→ only (most) efficient for large samples + one heuristic	
Best-first samples	→ best for small sample size + most common heuristics	
	→ can be drastically worse than other sample types in certain scenarios	
Inv-HS-Tree samples	→ best for medium sample size + two heuristics	
Larger samples	→ better estimations, but no higher diagnostic efficiency	(in general)
Better estimates	→ no higher diagnostic efficiency	(in general)
Time-information trade-off in diagnostic sampling		(most efficient sampling does not yield most effective samples, and vice versa)