Spectral debugging: How much better can we do?

Lee Naish
Hua Jie (Jason) Lee
Kotagiri Ramamohanarao

Computing and Information Systems
University of Melbourne

Slides, paper etc. are on the web:

http://www.cs.mu.oz.au/˜lee/papers/relscr/

Outline

Background: Spectral debugging

Rational metrics

Measuring performance

Single-bug optimality

Experimental results

Conclusion

Using spectra for bug localization

Basic idea:

Execute the program multiple times using a test suite where we can tell if each result is correct or not, gathering data about each execution

For each statement/. . . , estimate how likely it is to be buggy based on the data gathered

Rank the statements accordingly, then check the code manually, starting with the highest ranked statement until the bug is found (or we give up)

Statement spectra

Collect data on which statements are executed for each test

We count

- The total number of failed tests, F,
- The total number of passed tests, P,

and for each statement S_i, the number of

- failed tests in which it was executed, a^i_{ef}, and
- passed tests in which it was executed, a^i_{ep}.

(the number of failed/passed tests *not* executing S_i is implicit in our presentation)

Statement spectra example

The raw data is a binary matrix (1 means the statement was executed in the test) and a binary vector (1 means the test failed)

We compute F, P and the a_{ef} and a_{ep} for each statement, eg

	T_1	T_2	T_3	T_4	T_5	a_{ef}	a_{ep}
S$_1$	1	0	0	1	0	1	1
S$_2$	1	1	0	1	0	2	1
S$_3$	1	1	1	0	1	2	2

$\vdots$

Res.	1	1	0	0	0

$F = 2 \quad P = 3$

Measure “similarity” of matrix rows and result vector

Some metrics used for ranking

Many similarity metrics have been used, in a wide variety of domains, eg

Name	Formula	Name	Formula
Jaccard	$\frac{a_{ef}}{F+a_{ep}}$	Tarantula	$\frac{\frac{a_{ef}}{F}}{\frac{a_{ef}}{F}+\frac{a_{ep}}{P}}$
Russell	$\frac{a_{ef}}{F+P}$	Zoltar	$\frac{a_{ef}}{F+a_{ep}+\frac{10000F-a_{ef}*a_{ep}}{a_{ef}}}$
Ample	$\left\lvert\frac{a_{ef}}{F}-\frac{a_{ep}}{P}\right\rvert$	Ochiai	$\frac{a_{ef}}{\sqrt{F*(a_{ef}+a_{ep})}}$
O^p	$a_{ef}-\frac{a_{ep}}{P+1}$	Kulczynski2	$\frac{1}{2}\left(\frac{a_{ef}}{F}+\frac{a_{ef}}{a_{ef}+a_{ep}}\right)$

O^p performs best in our single-bug experiments and has been proven optimal in a very restricted setting

Kulczynski2 performs best in our multiple-bug experiments

Rational ranking metrics

A *(ranking) metric* is a partial function from four natural numbers, a_{ef}, a_{ep}, F and P to a real number. It is undefined if $a_{ef} > F$ or $a_{ep} > P$

Metrics measure *similarity*

A metric M is *rational* if it is monotonically increasing in a_{ef} and monotonically decreasing in a_{ep}: if $a'_{ef} > a_{ef}$ then $M(a'_{ef}, a_{ep}, F, P) \geq M(a_{ef}, a_{ep}, F, P)$ and if $a'_{ep} > a_{ep}$ then $M(a_{ef}, a'_{ep}, F, P) \leq M(a_{ef}, a_{ep}, F, P)$, for points where M is defined

We define *strictly rational* metrics similarly, using strict inequalities

Ample is the only proposed metric we know of which is not rational; it performs very poorly overall

Russell is rational but not strictly rational; if we tweek it to become strictly rational it is equivlent to O^p and performs better

Measuring performance — rank cost

Given a ranking of S statements, the *rank cost* is

$$\frac{GT + EQ/2}{S}$$

where GT is the number of correct statements ranked strictly higher than all bugs and EQ is the number of correct statements ranked equal to the highest ranked bug

This is similar to measures used by others

Ties in the ranking are handled in various ways; we think our approach is reasonable and it simplifies various things

Partial order for statements

The spectra associated with statements (plus rationality considerations) gives rise to a natural partial order

For two statements, x and y, with associated spectra:

- $x \leq^s y$ if $a^x_{ef} \leq a^y_{ef} \ \wedge \ a^x_{ep} \geq a^y_{ep}$
- $x =^s y$ if $a^x_{ef} = a^y_{ef} \ \wedge \ a^x_{ep} = a^y_{ep}$
- $x <^s y$ if $x \leq^s y \ \wedge \ \neg(a^x_{ef} =^s a^y_{ef})$

If $x \leq^s y$, any rational metric will rank x below or equal to y (rational metrics lead to a total order which is compatible with the partial order)

If $x <^s y$, any strictly rational metric will rank x below y

Unavoidable cost

Given a set of S statements and corresponding spectra, the *unavoidable cost* is the minimum of UC_b, for all bugs b, where

- $UC_b = \frac{GT'_b + EQ'_b/2}{S}$
- GT'_b is the number of correct statements c, such that $b <^s c$,
- EQ'_b, the number of correct statements c, such that $b =^s c$

For any set of statements and associated spectra, the unavoidable cost is the minimum rank cost for any strictly rational metric

Its clear the unavoidable cost is less than or equal to the rank cost for any strictly rational metric

With perfect knowledge, its possible to construct a strictly rational metric which has exactly the unavoidable cost

Unavoidable cost (cont.)

Let b be a bug which minimises UC_b

We can define a strictly rational metric M which achieves the unavoidable cost as follows:

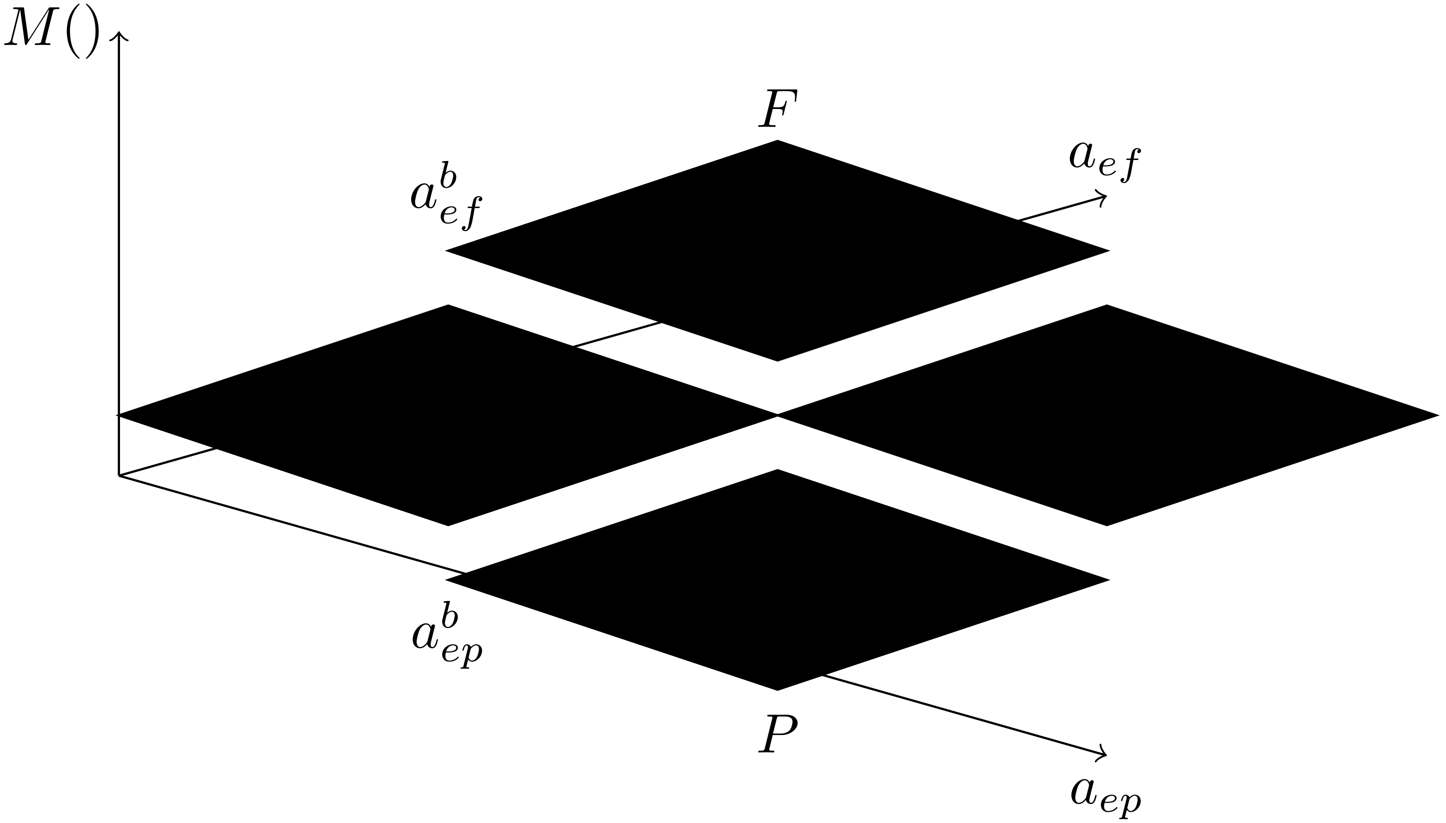
$$M(a_{ef}, a_{ep}, F, P) = f(a_{ef} - a^b_{ef}) + f(a^b_{ep} - a_{ep})$$

$$f(x) = \begin{cases} \epsilon x & \text{if } x < 0 \\ 1 + \epsilon x & \text{otherwise,} \end{cases}$$

$$\epsilon = 1/(F + P + 1)$$

The value for a statement c is 2 iff $b =^s c$ (unavoidable ties) and greater than 2 iff $b <^s c$

Unavoidable cost (cont.)



Unavoidable cost (cont.)

Metrics which are not strictly rational can result in less than the unavoidable cost

For example, the metric can return a very high value when $a_{ef} = a^b_{ef}$ and $a_{ep} = a^b_{ep}$

But we have no way of knowing a priori which metrics will perform the best (we don't know a^b_{ef} and a^b_{ep} if we don't know the buggy statement b)

Such metrics will typically perform poorly for other programs and spectra

Single-bug optimality

A metric M is *single bug optimal* if

1. when $a_{ef} < F$, the value returned is always less than any value returned when $a_{ef} = F$, that is, $\forall F \forall P \forall a_{ep} \forall a'_{ep}$ if $a_{ef} < F$ then $M(a_{ef}, a_{ep}, F, P) < M(F, a'_{ep}, F, P)$, and
2. when $a_{ef} = F$, M is strictly decreasing in a_{ep}, that is, if $a'_{ep} > a_{ep}$ then $M(F, a'_{ep}, F, P) < M(F, a_{ep}, F, P)$.

Optimality was shown previously for a single “model” program (with just four statements), using a simplistic performance measure (based on how often the bug was ranked top or equal-top) and assuming a particular distribution of sets of sets of tests

There is also empirical evidence such metrics perform best with single-bug programs

Single-bug optimality (cont.)

Given any program with a single bug, any set of test cases and any single bug optimal metric M used to rank the statements, the rank cost equals the unavoidable cost

Thus, the cost is no more than the rank cost using any other strictly rational metric

The same applies with most other reasonable measures of cost

The only previously known cases of optimal metrics being out-performed for single-bug programs are with the Russell metric, which benefits from ties in these cases

Optimizing metrics for single bugs

Any metric can be tweeked so it is single-bug optimal

The *optimal single bug version* of a metric M, denoted $O1(M)$ is defined as follows

$$O1(M)(a_{ef}, a_{ep}, F, P) = \begin{cases} K + 1 + P - a_{ep} & \text{if } a_{ef} = F \\ M(a_{ef}, a_{ep}, F, P) & \text{otherwise,} \end{cases}$$

where K is the maximum of $\{M(x, y, F, P) | x < F \ \wedge \ y \leq P\}$

In practice we can just use (eg) K=999999

If $a_{ef} < F$ for a large proportion of statements, $O1(M)$ will produce a similar ranking to M

Experimental results

Benchmark	1 Bug	2 Bug	2 Bug
Unavoidable	16.87	11.72	$O1(\ldots)$
O^p	16.87	21.64	21.64
Wong3	17.20	21.34	21.56
Zoltar	17.24	19.32	21.42
Kulczynski2	18.07	18.32	21.24
Ochiai	20.63	18.95	21.18
Jaccard	22.65	19.87	21.20
Tarantula	26.10	21.91	21.54
Ample	29.17	23.26	21.75
Russell	29.02	30.88	21.82

Experimental results (cont.)

To understand relative performance we looked at cases where $a_{ef} = F$ in more detail, and compared O^p and Kulczynski2 rankings

$a_{ef} = F$ for ...	2 Bugs	1 Bug same	1 Bug inverted	No Bug
% of Cases	44	35	11	10
% $a_{ef} = F$	67	51	44	37
Unavoidable	17.55	9.53	2.43	3.25
Kulczynski2	18.62	21.91	4.52	17.80
O1(Kulczynski2)	17.55	20.88	17.51	42.43
O^p	17.55	20.88	17.51	46.50
Russell	32.96	25.40	21.83	48.40

Experimental results (cont.)

Breakdown according to percentage of statements executed in all failed tests:

% $a_{ef} = F$	<20	20–40	40–60	60–80	≥80
% of Cases	10.6	8.1	19.0	57.9	4.2
Kul2	7.05	10.99	17.67	21.33	19.65
O1(Kul2)	5.47	16.08	21.51	24.51	21.89
O^p	6.24	19.17	21.64	24.58	21.75
O	16.70	23.39	23.57	24.96	22.37
Russell	8.35	22.59	26.11	36.26	43.96

Conclusions

Restricting attention to strictly rational metrics seems reasonable from a philosophical and empirical perspective

It allows us to give a much stronger result for single-bug optimal metrics — for single-bug programs we can't improve performance

All metrics can be adapted so they are single-bug optimal

Performance of single-bug optimal metrics on our multiple-bug benchmarks is adversely affected by the large proportion of statements which are executed in all failed tests

This proportion is smaller for larger benchmarks and potentially could be reduced by different test selection strategies

There are reasonable prospects for improving performance when there is a mixture of single- and multiple-bug programs