Test Set Diameter and Test Set Diversity

Robert Feldt, Simon Poulding, **David Clark** and Shin Yoo

Test Set Diameter

Idea

Use information theory to measure test set diversity

- Diversity $\longleftrightarrow$ Randomness $\longleftrightarrow$ Information
- Generic
- Universal

Universal, Generic, Similarity Metric

- Don't know probability distribution on inputs
- Test Set $\longleftrightarrow$ Set of Objects
- Minimise the similarity between the Objects
- Normalised Compression Distance for multisets (NCD) applied to test inputs (Input TSDm)

What is NCD for multisets?

$K(x|y)$: The conditional Kolmogorov complexity

The conditional Kolmogorov complexity of a string of symbols, x, given another string, y, is the length of the shortest program that outputs x, given y as input.

$K(x) \stackrel{\triangle}{=} K(x|\epsilon)$. Not computable.

ID: The Information Distance

For two strings x and y,

$$ID(x, y) = \max\{K(x|y), K(y|x)\}$$

Universal and Generic

NID: The Normalised Information Distance

For two strings x and y,

$$\mathrm{NID}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$

Enables comparisons between strings of different lengths

NCD: The Normalised Compression Distance

For two strings x and y,

$$\mathrm{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Computable approximation using compressors such as 7zip, Bzip

Extension to Multisets

Cohen and Vitanyi. *Normalised compression distance of multistep with applications*. 2012.

NCD: The Normalised Compression Distance for multisets

For a multiset X,

$$\mathrm{NCD}_1(X) = \frac{C(X) - \min_{x \in X}\{C(x)\}}{\max_{x \in X}\{C(X \setminus \{x\})\}} \tag{1}$$

$$\mathrm{NCD}(X) = \max\left\{\mathrm{NCD}_1(X), \max_{Y \subset X}\{\mathrm{NCD}(Y)\}\right\} \tag{2}$$

Time complexity $\mathcal{O}(2^{|X|})$

Approximate NCD for multisets

The algorithm starts from the multiset $Y_0 = X = \{x_1, x_2, \ldots, x_n\}$, and proceeds as:

1. Find index i that maximizes $C(Y_k \setminus \{x_i\})$.
2. Let $Y_{k+1} = Y_k \setminus x_i$.
3. Repeat from step 1 until the subset contains only two strings.
4. Calculate NCD(X) as: $\max_{0 \leq k \leq n-2}\{\text{NCD}_1(Y_k)\}$.

Time complexity $\mathcal{O}|X|^2)$

The nature of diversity

"Select diverse test cases"

- Inputs
- Outputs
- Execution Traces
- Combinations of these.

I-TSDm could be used before implementation with only partial specification or description

Research Questions

RQ1 **Correlation to code coverage**: Are higher levels of I-TSDm associated with higher levels of code coverage?

RQ2 **Structural coverage ability**: Do test sets selected based on I-TSDm lead to higher code coverage than randomly selected test sets?

RQ3 **Structural coverage ability w. size constraints**: Do test sets selected based on I-TSDm lead to higher code coverage than randomly selected test sets when we control for the size of test inputs?

RQ4 **Fault finding ability**: Do test sets selected based on I-TSDm lead to higher fault coverage than test sets based on random selection?

RQ5 **Selection time**: How does the time to execute the selection method scale as the size of the initial pool increase?

SUTs

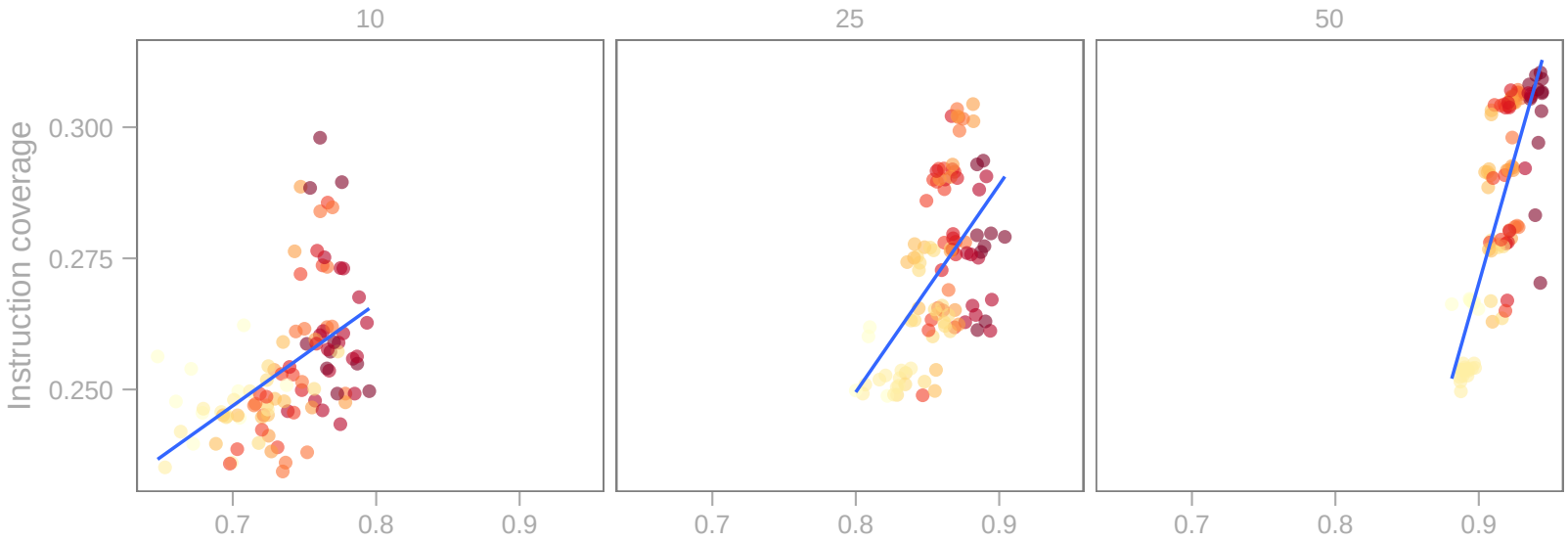
- **JEuclid** 3.1.9 Java library that renders images from MathML, an XML format for describing the presentation of mathematical equations. 11,556 SLOC.
- **ROME** 1.0 Java library for parsing and converting RSS and Atom formats for syndication feeds. 11,704 SLOC.
- **NanoXML** Small Java library for parsing XML. 1,603 SLOC.
- **Replace** C application from Software-artifact Infrastructure Repository. Takes three inputs: a string to be modified, a regular expression that defines matching text, and a string that replaces the matching text. 538 SLOC.

Correlation to Code Coverage

	Test Set Size		
SUT	**10**	**25**	**50**
JEuclid	0.59	0.67	0.52
NanoXML	0.50	0.40	0.26
ROME	0.60	0.57	0.82

- Used JaCoCo 0.7.4 to measure code coverage
- Sample from an initial pool of randomly generated test cases
- Vary the size of the subset (sensitivity to test set size)
- Calculated Spearman rank correlation at p-values less than 10^{-4}
- Moderate positive correlation

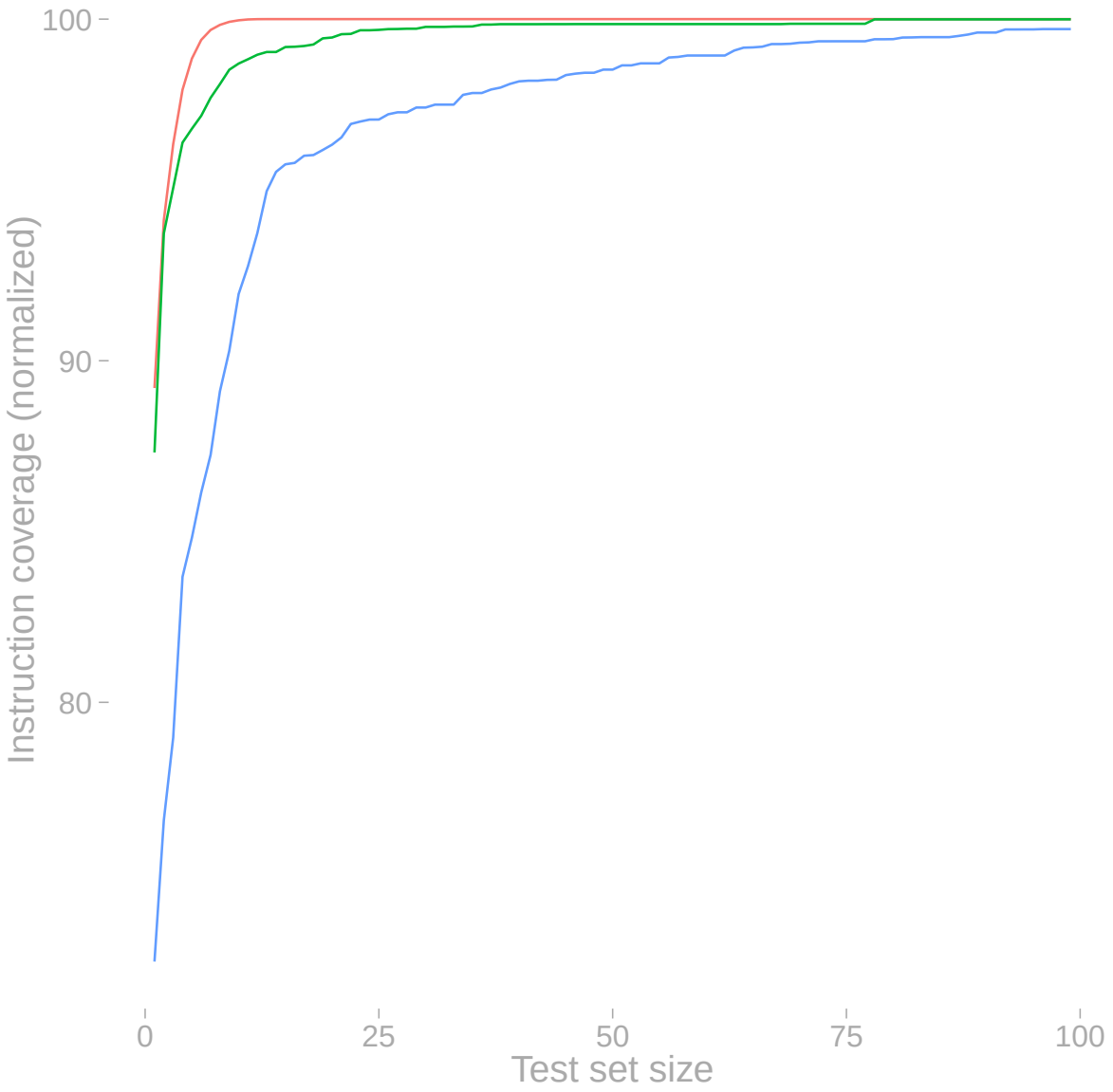
Code Coverage for increasing Test Set Size (ROME)



Test sets selected for highest test input diameter lead to higher code coverage than randomly selected test sets.

Structural Coverage

- Compare to a random selection of test sets (historical base point – does not need implementation)
- Compare to the post-hoc greedy algorithm (upper limit on what is achievable)
- Test inputs selected from an initial pool of 250 inputs


100
90
80
Instruction coverage (normalized)
0
25
50
75
100
Test set size

	Avg. Test Set Size					
	I-TSDm			Random		
SUT	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**
JEuclid	29.9	40.9	90.3	82.2	135.3	217.3
NanoXML	1.9	19.4	75.1	18.7	38.2	207.2
ROME	9.1	21.7	51.3	21.9	51.0	129.0

Average test set size needed to reach 90%, 95%, and 99% of the maximum instruction coverage reached by the greedy algorithm when selecting test inputs using the I-TSDm$_1$ procedure and the random algorithm from an initial pool of 250 inputs with lengths between 90 and 110 bytes

Test sets selected for highest test input diameter lead to higher code coverage than randomly selected test sets.

Other experiments

- Test set diameter can lead to higher code coverage even if we control for the size of test inputs; test diversity is more than simply the input length.
- Test sets with larger test set diameter (I-TSDm) may have better fault-finding ability.
- The TSDm test selection procedure scales quadratically in the size of the initial pool of tests to select from, and linearly with the average length of the tests.