How to read papers

Prerequisite for researchers:
You really should like to read.

I'm not talking about reading papers. I mean just reading. Anything.

I'd rather have a book, but in a pinch I will settle for a set of Water Pik instructions. Once, long ago, I bested a desperate bout of insomnia by studying the only piece of written material in my apartment that I had not already read at least twice: my roommate's 1974 Toyota Corolla manual.

– Ex Libris: Confessions of a Common Reader ("서재 결혼시키기"), Anne Faldman

Anatomy of Research Papers

- ✗ Title
- ✗ Authors
- ✗ Abstract
- ✗ Introduction
- ✗ Backgrounds (or Related Work)
- ✗ Technique X
- ✗ Research Questions
- ✗ Experimental Setup
- ✗ Results
- ✗ Discussions / Future Work
- ✗ Threats to Validity
- ✗ Related Work (or Backgrounds)
- ✗ Conclusion
- ✗ Acknowledgement
- ✗ References
- ✗ Appendix
- ✗ Online Supplements

Reading vs. Writing

- Much of what we discuss today also work as advice for writing (i.e., write what where)
- One helpful meta-reading is to focus on writing skills rather than contents – hard, but worthy skill to learn
  - Make note of words, phrases, even context-specific references

Title

“A rose by any other name would smell as sweet”
– Romeo & Juliet, Act 2 Scene 1
But really? ☺

Title

- ✗ What the paper is about, obviously
- ✗ The "granularity"
  - "On [something really big]"
  - "Towards [something really difficult]"
  - "Evaluating the impact of X-zation of Y-ified Z on [something widely studied]"

Authors

“If my doctor told me I had only six minutes to live, I would not brood. I’d type a little faster.”
– Isaac Asimov

Authors

- ✗ Knowing individual authors do help you reading the paper
  - Research trajectory
  - Area of expertise
  - Expected quality (ideally)
  - Social landscape of a field
- ✗ Remembering the mapping between names and papers will help you when you actually run into them at conferences ☺

Abstract

"Yes."
– J. K. Gardner & L. Knopoff

Abstract

- ✗ A short summary of the entire paper
- ✗ Usually 250–300 words
- ✗ Some venues require structured abstract
  - Introduction / Methods / Results / Conclusion (IMRaD)
- ✗ You SHOULD get some idea about what the paper is really about

JOURNAL ARTICLE

Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?

J. K. Gardner; L. Knopoff

Bulletin of the Seismological Society of America (1974) 64 (5): 1363-1367.
Published: October 01, 1974 Article history ▾

Cite Share ▾ Tools ▾

Abstract

Yes.

Introduction

"You never get a second chance to make the first impression."
– anonymous

Introduction

- ✗ A slightly longer summary of the entire paper (abstract expanded)
- ✗ Ordering can be different, but usually contains the following parts:
  - Description of the world/status-quo
    - What the problem is
    - Why the problem is important (i.e., motivation)
  - What other people have tried before
    - Why they fall short
  - Description of the current paper
    - What "we" do for the problem, differently from others (i.e., justification)
  - A brief summary of technical contributions
  - Organisation of the remaining sections

2016 IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops

Diversity-Aware Mutation Adequacy Criterion for Improving Fault Detection Capability

Donghwan Shin
School of Computing
KAIST
Daejeon, Republic of Korea
Email: donghwan@se.kaist.ac.kr

Shin Yoo
School of Computing
KAIST
Daejeon, Republic of Korea
Email: shin.yoo@kaist.ac.kr

Doo-Hwan Bae
School of Computing
KAIST
Daejeon, Republic of Korea
Email: bae@se.kaist.ac.kr

***Abstract*—Many existing testing techniques adopt diversity as an important criterion for the selection and prioritization of tests. However, mutation adequacy has been content with simply maximizing the number of mutants that have been killed. We propose a novel mutation adequacy criterion that considers the diversity in the relationship between tests and mutants, as well as whether mutants are killed. Intuitively, the proposed criterion is based on the notion that mutants can be distinguished by the sets of tests that kill them. A test suite is deemed adequate by our criterion if the test suite distinguishes all mutants in terms of their kill patterns. Our hypothesis is that, simply by using a stronger adequacy criterion, it is possible to improve fault detection capabilities of mutation-adequate test suites. The empirical evaluation selects tests for real world applications using the proposed mutation adequacy criterion to test our hypothesis. The results show that, for real world faults, test suites adequate to our criterion can increase the fault detection success rate by up to 76.8 percentage points compared to test suites adequate to the traditional criterion.**

I. Introduction

One fundamental limitation of software testing is the fact that, to validate the behavior of the Program Under Test (PUT), we can only ever sample a very small number of test inputs out of the vast input space. Almost all existing testing techniques are, at some level, attempts to answer the following question: *how does one sample a finite number of test inputs to cover a wide a range of program behaviours as possible?*

The concept of diversity has received much attention while answering the above question. For example, Adaptive Random Testing (ART) [1] seeks to increase diversity of randomly sampled test inputs by choosing an input that is as different from those already sampled as possible. Clustering-based test selection and prioritization [2], [3] assumes that a diverse set of test inputs would explore and validate a wider range of program behaviors. Diversity in test output has been studied as a test adequacy criterion for black box testing of web applications [4]. Information theoretic measures of diversity has also been studied as a test selection criterion [5], [6].

In contrast, diversity has received little attention in relation to mutation testing. Mutation adequacy remains essentially as a simple count of the number of killed mutants. Many of existing works focus either on reducing the cost of mutation testing (i.e., *do fewer, do smarter, and do faster* as summarized by Offutt and Untch [7]) or analyzing equivalent mutants [8], [9] (i.e., mutants semantically equivalent to an original program). Relatively very few attention has been paid to improve the fault detection capability of the mutation adequacy criterion itself.

The existing mutation adequacy as a count of killed mutants does not cater for diversity. Consequently, despite its potential correlation to the fault detection capability, many diverse mutants are generated but wasted. Suppose a pathological case in which a single test can kill all generated mutants. In terms of the kill information, it means that the single test does not capture the diversity of the mutants enough, while the traditional mutation adequacy simply determines the single test as adequate. Such a case calls for a richer adequacy criterion in mutation testing.

This paper proposes a novel adequacy criterion called distinguishing mutation adequacy criterion, which includes the notion of diversity. The proposed metric is based on our previous work on theoretical framework for mutation testing [10]. At the core of the new criterion lies the idea that mutants can be *distinguished* from each other by the set of tests that kill them. Our mutation adequacy criterion aims not only to kill, but also to distinguish as many mutants as possible. The aforementioned pathological case of a single test killing all mutants will perform poorly under our new criterion.

By aiming to distinguish the maximum number of mutants, the proposed adequacy criterion can select more diverse set of test cases. Suppose there exist two mutants. Test t_1 kills both, while t_2 and t_3 kill different one each. Under the existing criterion, the set $\{t_1\}$ is deemed adequate, whereas the proposed criterion will choose the set $\{t_2, t_3\}$ instead. We hypothesize that this, more diverse set of tests will show higher fault detection capability.

The hypothesis on fault detection capability is validated by an empirical study. We use the `Defects4J` [11] data set to study real world faults in non-trivial systems. The control group consists of sets of test suites selected based on the traditional mutation adequacy, i.e., ones that kill all mutants, while the treatment group consists of sets of test suites that can collectively kill and distinguish all mutants. Both groups are evaluated for their fault detection capabilities by executing them against faulty and fixed versions of programs collected from the real world. The results show that our novel mutation

978-1-5090-3674-5/16 $31.00 © 2016 IEEE
DOI 10.1109/ICSTW.2016.37

adequacy criterion shows either equal or higher fault detection capabilities for all studied subject faults.

The technical contributions of this paper are as follows:

- This paper introduces a novel mutation adequacy criterion called *distinguishing mutation criterion*. A test suite is mutation adequate with respect to this criterion if all considered mutants have unique sets of tests that kill them, i.e. can be *distinguished* by their kill patterns.
- The proposed adequacy criterion is empirically evaluated using real faults in the `Defects4J` repository and random testing. The results show that the new adequacy criterion shows at least equal or higher fault detection capability than the traditional mutation adequacy criterion. The increase in the fault detection success rate is up to

The rest of the paper is organised as follows. Section II presents a formal definition of the existing mutation adequacy criterion. Section III introduces the distinguishing mutant adequacy criterion using the same formal notations. Section IV describes the design of our empirical evaluation, the results of which are presented and analysed in Section V. Section VI discusses related work, and finally Section VII concludes.

II. Background

A. Formal Model of Mutation Testing

To formally represent mutation adequacy criteria considered in this paper, we summarize the essential elements of the formal framework for the mutation-based testing methods. Detailed descriptions for the formal framework are presented in [10].

Let P be a set of programs which includes the program under test. In mutation testing, there are three essential programs in P: an original program $p_o \in P$, a mutant $m \in M \subseteq P$ generated from p_o, and a correct program $p_c \in P$ which represents the true requirements[1] about p_o. For a test $t \in T$ for P, if the behaviors of p_c and p_o are different, it is said that t *detects* a fault in p_o. Similarly, if the behaviors of p_o and m are different for t, it is said that t *kills* m. Note that the notion of behavioral difference is an abstract concept. It is formalized by a testing factor, called a test *differentiator*, which is defined as follows:

Definition 1: A *test differentiator* $d : T \times P \times P \rightarrow \{0, 1\}$ is a function,[2] such that

$$d(t, p_c, p_y) = \begin{cases} 1\ (true), & \text{if } p_y \text{ is } \textit{different} \text{ with } p_y \text{ for } t \\ 0\ (false), & \text{otherwise} \end{cases}$$

for all tests $t \in T$ and programs $p_x, p_y \in P$.

By definition, a test differentiator concisely represents whether the behaviors of $p_y \in P$ and $p_y \in P$ are different for t.

[1]While p_c is not a real program, this is not a serious assumption, because we only require the behavior of p_c for a given set of tests. In practice, a human may play the role of p_c, acting as a human oracle.

[2]This function-style definition is replaceable by a predicate-style definition, such as $d \subseteq T \times P \times P$.

We make no attempt to incorporate any specific definition of program differences. The specific definition of differences can only be decided in context. For example, while 0.3333 is different with 1/3 in the strict sense, 0.3333 will be regarded as the same as 1/3 in some cases. To keep things general, we consider a set of test differentiators D that includes all possible test differentiators for P.

A test differentiator, or simply differentiator, can formally describe the notion of differences in mutation testing. For example, when t detects a fault in p_o, it is clearly formalized as follows[3]:

$$d(t, p_c, p_s) = 1$$

On the other hand, when t kills a mutant m, it is also clearly formalized as follows:

$$d(t, p_s, m) = 1$$

Note that p_c, p_s, and m are general entities, and largely separated from any specifics such as programming languages or mutation methods.

B. Mutation Adequacy Criterion

Since mutation testing was first proposed in the 1970s, it has been widely studied in the aspects of both theory and practice, and a mutation adequacy criterion has played the key role in the studies of mutation testing. A mutation adequacy criterion is a predicate that determines the adequacy of a test suite using mutants. It is said that a test suite is *mutation-adequate* when the test suite kills all of the generated mutants. Using a differentiator, it is clearly and concisely formalized as follows:

$$\forall m \in M, \exists t \in TS, d(t, p_o, m) = 1. \qquad (1)$$

In other words, a test suite TS is mutation-adequacy if all mutants $m \in M$ are killed by at least one test $t \in TS$.

Equation (1) is general enough to consider various mutation testing approaches. For example, there is a spectrum of mutation approaches from a strong mutation [12] to a weak mutation [13], depending on which d is used. In a strong mutation analysis, a test t kills a mutant m when the output of m differs from the output of the original program p_o for t. In a weak mutation analysis, t kills m when the internal states of m and p_o are different for t. In the rest of this paper, we refer (1) as the traditional mutation adequacy criterion in compared to the new mutation adequacy proposed in Section III-B.

III. Extending Mutation Adequacy Considering Diversity of Mutants

A. Limitation of Traditional Mutation Adequacy

To see the limitation of the traditional mutation adequacy criterion, we provide a working example with four mutants and three tests in Figure 1. Each of the values represents whether a test kills a mutant. For example, $d(t_1, p_o, m_3)$ is 1 which means that t_1 kills m_1.

[3]In experiments, when the correct version of a program for a fault is known in advance, the correct version can be used as p_c. In this case, the corresponding faulty version should be used as p_s so that the difference between p_c and p_s implies the fault.

Part 1: State of the world

One fundamental limitation of software testing is the fact that, to validate the behavior of the Program Under Test (PUT), we can only ever sample a very small number of test inputs out of the vast input space. Almost all existing testing techniques are, at some level, attempts to answer the following question: *how does one sample a finite number of test inputs to cover as wide a range of program behaviours as possible?*

The concept of diversity has received much attention while answering the above question. For example, Adaptive Random Testing (ART) [1] seeks to increase diversity of randomly sampled test inputs by choosing an input that is as different from those already sampled as possible. Clustering-based test selection and prioritization [2], [3] assumes that a diverse set of test inputs would explore and validate a wider range of program behaviors. Diversity in test output has been studied as a test adequacy criterion for black box testing of web applications [4]. Information theoretic measures of diversity has also been studied as a test selection criterion [5], [6].

- ✗ State of the fact
- ✗ What has been happening in other field

Part 2: Description of the Problem

In contrast, diversity has received little attention in relation to mutation testing. Mutation adequacy remains essentially as a simple count of the number of killed mutants. Many of existing wo (i.e., mutants semantically equivalent to an original program). (i.e Relatively very few attention has been paid to improve the Off fault detection capability of the mutation adequacy criterion itself.

The existing mutation adequacy as a count of killed mutants does not cater for diversity. Consequently, despite its potential correlation to the fault detection capability, many diverse mutants are generated but wasted. Suppose a pathological case in which a single test can kill all generated mutants. In terms of the kill information, it means that the single test does not capture the diversity of the mutants enough, while the traditional mutation adequacy simply determines the single test as adequate. Such a case calls for a richer adequacy criterion in mutation testing.

- ✗ Compared to what has been going on in other fields, ours hasn't put much effort in X
- ✗ Existing Y fails to do X because...

Part 3: What we propose

This paper proposes a novel adequacy criterion called distinguishing mutation adequacy criterion, which includes the notion of diversity. The proposed metric is based on our previous work on theoretical framework for mutation testing [10]. At the core of the new criterion lies the idea that mutants can be *distinguished* from each other by the set of tests that kill them. Our mutation adequacy criterion aims not only to kill, but also to distinguish as many mutants as possible. The aforementioned pathological case of a single test killing all mutants will perform poorly under our new criterion.

By aiming to distinguish the maximum number of mutants, the proposed adequacy criterion can select more diverse set of test cases. Suppose there exist two mutants. Test t_1 kills both, while t_2 and t_3 kill different one each. Under the existing criterion, the set $\{t_1\}$ is deemed adequate, whereas the proposed criterion will choose the set $\{t_2, t_3\}$ instead. We hypothesize that this, more diverse set of tests will show higher fault detection capability.

- ✗ High level description of what this paper proposes
- ✗ Try to imagine what the low-level, technical details would be

Part 4: How we are going to evaluate

The hypothesis on fault detection capability is validated by an empirical study. We use the `Defects4J` [11] data set to study real world faults in non-trivial systems. The control group consists of sets of test suites selected based on the traditional mutation adequacy (i.e., ones that kill all mutants), while the treatment group consists of sets of test suites that can collectively kill and distinguish all mutants. Both groups are evaluated for their fault detection capabilities by executing them against faulty and fixed versions of programs collected from the real world. The results show that our novel mutation

- ✗ Peek into the actual evaluation
- ✗ Allows you to set your expectation – what evidence will I see?

Part 5 & 6: Contributions and Organisation

The technical contributions of this paper are as follows:

- This paper introduces a novel mutation adequacy criterion called *distinguishing mutation criterion*. A test suite is mutation adequate with respect to this criterion if all considered mutants have unique sets of tests that kill them, i.e. can be *distinguished* by their kill patterns.
- The proposed adequacy criterion is empirically evaluated using real faults in the `Defects4J` repository and random testing. The results show that the new adequacy

The rest of the paper is organised as follows. Section II presents a formal definition of the existing mutation adequacy criterion. Section III introduces the distinguishing mutants adequacy criterion using the same formal notations. Section IV describes the design of our empirical evaluation, the results of which are presented and analysed in Section V. Section VI discusses related work, and finally Section VII concludes.

- ✗ Point- by-point, declarative summary of what the paper actually contributes
- ✗ A brief ToC...

Background / Related Work

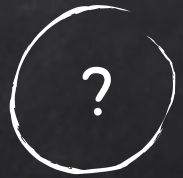
"Knowledge is power."
- Francis Bacon
(Bonus: definitely read this AFTER CLASS- https://www.reddit.com/r/AskReddit/comments/dxosj/what_word_or_phrase_did_you_totally_misunderstand/c13pbyc/)

Background / Related Work

- ✗ Where you list existing work that are related to yours
- ✗ Depending on context, it may be better to:
  - Explain existing material before describing yours: have a “Background” section
  - Describe your technique first, and compare it to existing work before: have a “Related Work” section
  - Can have both in rare cases
- ✗ Summarise existing work briefly
- ✗ *Position your work*
  - highlight differences between them and yours

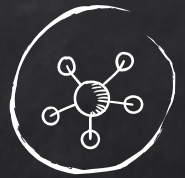
Research Questions / Experimental Setup

"Judge a man by his questions rather than his answers."
– Voltaire



Research Questions

- ✗ The “solution” in a research question is usually presented in the form of question and answer.
- ✗ You can learn how the proposed solution is being evaluated by reading RQs
  - Effectiveness: how good is the new technique?
  - Efficiency: how cheap/fast is the new technique?



Experimental Setup

- For domains that involves empirical evaluation (i.e., your "solution" is a "technique" therefore you apply it to a set of subjects and evaluate how well it works), you need to specify:
  - What your subjects are, and why/how they were chosen
  - How your technique is implemented, using which external dependencies
  - In which environment the experiments have been executed
  - What the parameters to your algorithm are
- This helps both replication and understanding of results

Threats to Validity

"The validity of internet quotes are getting sketchy these days."
– Abraham Lincoln

Threats to Validity

- ✗ This is where you honestly discuss any factors that may threat the validity of your claims
  - Internal validity: have you evaluated your solution correctly?
    - E.g., implementational correctness
  - External validity: how well does your claim generalise?
    - E.g., any potential bias in the selection of subjects
  - Construct validity: are you measuring what you are supposed to measure?
    - E.g., the use of correct evaluation metrics
- ✗ Use extra scepticism

Reading Strategy

It is OKAY to feel overwhelmed

- ✗ Reading research papers is difficult. By definition.
- ✗ Keep in mind *why* you are reading this paper (when applicable, ask the person who told you to read it)
  - Comfort zone vs. extension of your knowledge
  - Theoretical vs. empirical

Multi-pass Reading

- Almost all guidelines recommend multi-pass reading
  - Skim: focus on title, abstract, figures, and conclusion – identify what it is about
  - Read: read again, using the common structure as your map – form your own questions
  - Interpret: deep-dive into the text, tables, and figures – find answers to your own questions
  - Summarise: add the paper to your own map of research landscape

Critical/Creative Reading

- Read Critically
  - Do not be afraid to question the contents – be sceptical and suspicious
  - Is the problem right? Is the argument justifiable? Are the data not biased?
- Read Creatively
  - Criticising something is easy: building up is harder
  - Can this be done for other problems? What would I do differently, or subsequently?

Critical/Creative Reading

- ✗ Read Between/Beyond Lines
  - Notice what is NOT being said
  - Learn to distinguish nuances
- ✗ Reading as a group
  - Will bring out points that you missed
  - Discussion is not a fight to win or lose
  - Reading critically does not mean you have to **judge** the paper

"IT HAS LONG BEEN KNOWN"	I didn't look up the original reference.
"A DEFINITE TREND IS EVIDENT"	The data are practically meaningless.
"WHILE IT HAS NOT BEEN POSSIBLE TO PROVIDE DEFINITE ANSWERS TO THE QUESTIONS"	An unsuccessful experiment, but I still hope to get it published.
"THREE OF THE SAMPLES WERE CHOSEN FOR DETAILED STUDY"	The other results didn't make any sense.
"TYPICAL RESULTS ARE SHOWN"	This is the prettiest graph.
"THESE RESULTS WILL BE IN A SUBSEQUENT REPORT"	I might get around to this sometime, if published/funded.
"A CAREFUL ANALYSIS OF OBTAINED DATA"	Three pages of notes were obliterated when I knocked over a glass of beer.
"AFTER ADDITIONAL STUDY BY MY COLLEAGUES"	They didn't understand it, either.
"THANKS ARE DUE TO JOE BLOTZ FOR ASSISTANCE WITH THE EXPERIMENT AND TO CINDY ADAMS FOR VALUABLE DISCUSSIONS"	Mr. Blotz did the work and Ms. Adams explained to me what it meant.
"A HIGHLY SIGNIFICANT AREA FOR EXPLORATORY STUDY"	A totally useless topic selected by my committee.
"IN MY EXPERIENCE"	Once
"IN CASE AFTER CASE"	Twice
"IN A SERIES OF CASES"	Three times
"IT IS BELIEVED THAT"	I think.
"IT IS GENERALLY BELIEVED THAT"	A couple of others think so, too.
"CORRECT WITHIN AN ORDER OF MAGNITUDE"	Wrong.
"ACCORDING TO STATISTICAL ANALYSIS"	Rumor has it.
"IT IS CLEAR THAT MUCH ADDITIONAL WORK WILL BE REQUIRED BEFORE A COMPLETE UNDERSTANDING OF THIS PHENOMENON OCCURS"	I don't understand.
"A STATISTICALLY-ORIENTED PROJECTION OF THE SIGNIFICANCE OF THESE FINDINGS"	A wild guess.
"IT IS HOPED THAT THIS STUDY WILL STIMULATE FURTHER INVESTIGATIONS IN THIS FIELD"	I quit.

Both use and ignore the structure

- ✗ Different papers are written for different model readers in mind
  - Most frequently: for someone with general CS knowledge by little domain expertise
  - Depending on venues: for someone with reasonable domain knowledge on a specific topic (e.g., specialised conference proceedings)
- ✗ Do not read everything sequentially. Jump depending on:
  - Your domain knowledge
  - Your questions

Treat Citations as Hypertexts

- ✗ Be prepared to follow references and read other papers
- ✗ People *do* make habitual citations (I am guilty as well), sometimes completely out of context – do not rely on authority, find out on your own

Reproduction

- ✗ Increasingly researchers are required to provide a replication package with a paper (code, data, and anything else required to replicate the study)
- ✗ Make a note of whether such a package is explicitly included in the paper – it will save you time later

Read to Write Better

- ✗ Try to expand your vocabulary all the time: sometimes you want the prefect choice of word to get the right nuance
- ✗ Pick up phrases and expressions that are useful
- ✗ Identify fancy LaTeX tricks; maintain a snippet library

Note-making

- ✗ Write a short summary of what you read, if possible
- ✗ What the paper is about, which data has been used, what you think are strengths and weaknesses

Summaries

- ✗ There are common structures papers follow – know which contains which
- ✗ Do multi-pass reading, gradually digging into finer details
- ✗ Read both critically and creatively
- ✗ Maintain a note of what you read

More reading about reading

- How to read a paper – S. Keshav, University of Waterloo (link)
- How to read a research paper – M. Mitzenmacher, Harvard University (link)
- How to read a research paper – Grisword, Murphy, and Conati (link)
- How to (seriously) read a scientific paper – Science Magazine (link)
  - How to read a scientific paper (link)