On the Relationships between the Distribution of Failure-Causing Inputs and Effectiveness of Adaptive Random Testing

Tsong Yueh Chen                Fei-Ching Kuo+
Zhi Quan Zhou

Faculty of Information & Communication Technologies
Swinburne University of Technology
Hawthorn, 3122, Australia

Abstract
Recently, adaptive random testing (ART) has been developed to enhance the fault-detection effectiveness of random testing (RT). It has been known in generalities that the fault-detection effectiveness of ART depends on the distribution of failure-causing inputs, yet this understanding is in coarse terms without precise details. In this paper, we conduct an in-depth investigation into the factors that have an impact on the fault-detection effectiveness of ART. This paper gives a comprehensive analysis of the favourable conditions for ART and, hence, provides a guideline for testers to decide when to use ART instead of RT.

1. Introduction

Software testing is an important method of software quality assurance. One activity of software testing is to select effective test cases with a higher chance of revealing failures. A basic test case selection strategy is Random Testing (RT), which selects test cases in a random manner from the set of all possible inputs (known as the input domain) [12][17]. RT is good at revealing bugs which are usually overlooked by software development teams [11]. When other testing techniques are difficult to apply (probably due to the high costs involved in its application, or the lack of formal specifications or source code), RT can be a cost-effective method to automatically generate large quantities of test cases and to widely cover the input domain [9][20]. The above strengths make RT widely used in many real-life applications [8][9][11][14][15][16][18][20][22]. Despite the popularity of RT, some people consider it ineffective because RT does not use any specific information from the programs or their specifications to generate test cases.

It has been pointed out that failure-causing inputs tend to cluster together [1][2][10]. Chen et al. [7][13] have made use of this feature, that is, the distribution of failure-causing inputs, to improve the fault-detection effectiveness of RT. They found that if the random inputs are adaptively selected to be more evenly spread over the input domain, then program failures can be more effectively detected than with RT. They named this approach as Adaptive Random Testing (ART) [7][13]. A concern about the use of ART is the additional overhead involved in its test case generation process to evenly spread test cases. This is why, since the introduction of ART, research has been focused on how to minimize this overhead [5][6].

Because ART has been developed as an enhanced version of RT, it is natural to compare the fault-detection effectiveness of ART and RT. There is always a great interest to know under what conditions ART significantly outperforms RT, so that testers can decide when to use ART instead of RT. In this paper, we investigate the fault-detection effectiveness of ART under various scenarios, and report our findings that ART’s effectiveness is in fact closely correlated to several properties of the distribution of failure-causing inputs.

This paper is organized as follows. Section 2 provides some background information of ART. Section 3 describes several experiments and our findings. Section 4 analyzes the relationships between the distribution of failure-causing inputs and the fault-detection effectiveness of ART, and discusses the implication of our study.

2. Background

Any faulty program has two fundamental attributes: its failure rate (the ratio of the number of failure-causing inputs to the number of all possible inputs) and its failure pattern (the geometric shapes of the regions formed by the failure-causing inputs and the distribution of these regions within the input domain). Both attributes are fixed but unknown to testers in advance of testing. Chan et al. [4] coarsely classified failure patterns into three types, namely point, strip and block patterns. The point pattern refers to the situation where “failure-causing inputs are not concentrated in one or a few regions, but widely dispersed over a large part of the input domain”. The strip and block patterns refer to the situation where
failure-causing inputs form a narrow strip or a block region, respectively. These patterns are schematically depicted in Figure 1. Examples 1, 2 and 3 illustrate sample program faults giving rise to these patterns.

A. Point pattern         B. Strip pattern            C. Block pattern

**Figure 1. Three types of failure-causing patterns**

**Example 1: A program fault causing the point pattern**

```
INPUT X, Y
IF (X mod 2 = 0 AND Y mod 2 = 0)
    {  Z = X – Y    /* ERROR: Should be Z = X + Y */  }
ELSE
    {  Z = X*Y  }
OUTPUT Z
```

**Example 2: A program fault causing the strip pattern**

```
INPUT X, Y
IF (Y <= 0)    /* ERROR: Should be IF(Y < 0) */
    {  Z = X - Y  }
ELSE
    {  Z = X + Y  }
OUTPUT Z
```

**Example 3: A program fault causing the block pattern**

```
INPUT X, Y
IF (X >0 AND X <10 AND Y >0 AND Y <10)
    {  Z = X    /* ERROR: Should be Z = 2*X */  }
ELSE
    {  Z = 2*Y  }
OUTPUT Z
```

The Adaptive Random Testing (ART) method has been proposed to enhance the fault-detection effectiveness of Random Testing (RT) for the situations where failure-causing inputs are clustered together, as in block and strip patterns. In ART, test cases are not only randomly selected, but also evenly spread across the input domain. All the results of the previous experiments have shown that when the failure pattern of a program is not of point type, ART requires much fewer test cases than RT to detect the first failure (up to 50% saving) [7][13].

There are various implementations of ART. A simple approach is the Fixed Size Candidate Set version (FSCS-ART) [7][13]. In FSCS-ART, two sets of inputs are maintained, namely the executed set (E), and the candidate set (C). E stores test cases that have already been executed without revealing any failure, while C stores a number (in our experiments, 10) of randomly generated inputs, from which the next test case will be chosen. For each candidate \( c_i \) in C, its shortest distance to E (that is, the distance from \( c_i \) to the nearest element in E) is measured. The candidate with the maximum shortest distance will be selected as the next test case. Without ambiguity, ART refers to FCS-ART in the rest of our discussions, unless otherwise specified.

### 3. The Experiments

ART has been proposed to improve on the fault-detection effectiveness of RT. In this paper, we adopt the F-measure (that is, the number of test cases required to detect the first failure) to assess their effectiveness, and the ART F-ratio (that is, the ratio of ART’s F-measure (\( F_{ART} \)) to RT’s F-measure (\( F_{RT} \))) to indicate how much improvement ART has over RT. Obviously, the smaller the ART F-ratio is, the better the fault-detection effectiveness of ART is. Our study assumes that all inputs have the same probability of being chosen as test cases, and that selection is with replacement. The expected \( F_{RT} \) is known to be 1/\( \theta \), where \( \theta \) denotes the failure rate. The average \( F_{ART} \) was collected in experiments.

For a program with \( N \) input parameters, we say the input domain is of \( N \) dimensions. Our experiments have assumed that each dimension of the input domain has the same range of values (for example, a square when \( N=2 \) and a cube when \( N=3 \)). Adopting the same definition given in [1], failure region refers to the region occupied by the failure-causing inputs. For each run in the experiments, the failure regions were randomly placed, and once the first failure was detected, the total number of executed test cases was recorded as the \( F_{ART} \) of that run. The experiment was repeated \( S \) times to obtain a statistically significant mean value of \( F_{ART} \) with an accuracy range of \( \pm 5\% \), and a confidence level of 95%. Details of the method used to obtain \( S \) can be found in [6].

#### 3.1 Experiment 1

The aim of this experiment was to investigate the relationship between the ART F-ratio and the program failure rate (\( \theta \)), where \( \theta \) was varied from 0.00005 to 1. The failure pattern was set to be a single square located within the input domain, the number of dimensions of which was varied from 1 to 3. The results of the experiment are summarized in Figure 2.

This experiment showed that for the same \( \theta \), ART has a larger F-ratio for input domains of higher dimensions. Nevertheless, when \( \theta \) gets smaller, the difference between the ART F-ratios of different dimensional input domains becomes smaller. When \( \theta \) is very large from the perspective of testing (\( \theta > 0.75 \), \( \theta > 0.25 \) and \( \theta > 0.075 \) for 1D, 2D and 3D input domains, respectively), ART uses more test cases than RT to detect the first failure. This apparently surprising result is due to the fact that at the beginning FSCS-ART tends to select test cases around the corners or edges of the input domain. Test cases will be selected around the central part of the input domain only when a
sufficient number of test cases around the corners or edges have been selected (Please note that some other implementations of ART do not have this problem, for example, ART by bisection [5]). Hence, when \( \theta \) is large (that is, only a few test cases are needed to detect the first failure), FSCS-ART may perform worse than RT. This phenomenon becomes worse for the input domain of higher dimensions as the number of corners increases. We are currently working on an improved implementation of FSCS-ART to address this problem.

As can be seen in Figure 2, when \( \theta \) becomes smaller, ART F-ratio becomes smaller for input domains of all dimensions. We wish to emphasize that even when \( \theta \) is not small enough to give a significant saving in F-measure for ART, ART may still be more cost-effective than RT when the program executions and the verifications of computed outputs are expensive as compared with the test case generation.

Although some people consider RT an ineffective strategy when \( \theta \) is small, we would like to point out that in this situation ART can be an effective replacement of RT since ART preserves the simplicity nature of RT but uses significantly less test cases to detect the first failure at small \( \theta \) and, hence, brings in considerable cost savings.

There are many metrics to measure the compactness of a shape [19][23]. One of these measures [19][21] compares the shape’s size with the size of the circular shape having an equal boundary. According to this measurement, formulae (1) and (2) give the compactness ratio for 2D and 3D geometric shapes, respectively. The ratio ranges from 0 to 1, with the circle/ sphere having the largest value, 1, as it has been proved that a circle/ sphere encloses the largest region among all shapes having the same boundary [3].

\[
\frac{4\pi a^2}{P^2} \quad (1)
\]

\[
\frac{6\pi^2 a^3}{8P^3} \quad (2)
\]

In the formulae, \( A \) denotes the area of the circle or the volume of the sphere, respectively; \( P \) denotes the perimeter of the circle or the surface area of the sphere, respectively.

In Experiment 2, the edge lengths of the rectangular failure regions are in the ratio 1: \( r \) (2D) or 1: \( r/r \) (3D). It is straightforward to prove that a larger \( r \) indicates less compactness (proofs are given in the Appendix). The results of the experiment are shown in Figures 3 and 4. It is evident that the ART F-ratio increases when \( r \) increases, that is, when the failure region becomes less compact,
irrespective of $\theta$ and dimensions of the input domain. We wish to clarify that although failure regions in the experiments are rectangular, they are for illustration only. In reality, the shapes vary. Our ongoing experiments with other shapes of failure regions have also confirmed our hypothesis that different fault-detection effectiveness of ART for different shapes is due to their different compactness.

3.3 Experiment 3

It is known that ART has a significant improvement over RT when all failure-causing inputs are clustered in a block region, but no improvement when the failure-causing inputs are scattered over the input domain. However, the relationship between the number of failure regions and the ART F-ratio has never been precisely reported. Experiment 3 was designed to investigate this relationship. In this experiment, the number of failure regions ($n$) was varied from 1 to 100, and all these $n$ failure regions were square and of equal size. Both 2D and 3D input domains were investigated with $\theta$ being set to 0.005, 0.001 and 0.0005. The results are shown in Figures 5 and 6.

The result showed that for both 2D and 3D input domains, at the same $\theta$, the ART F-ratio increases with $n$ and reaches a plateau very quickly (around 20 and 10 failure regions for 2D and 3D input domains, respectively). This observation is in line with the intuition that ART favours the situations where failure-causing inputs are clustered together, because more failure regions mean weaker clustering of failure-causing inputs.

3.4 Experiment 4

Experiment 3 investigated how the number of failure regions impacts on the ART F-ratio with equally sized failure regions. In this section, we investigate the situations where the input domain contains failure regions of varied sizes.

Experiment 4 was conducted using a 2D input domain containing $n$ square failure regions, where $n$ was varied from 1 to 100, and $\theta$ was set to 0.001. We used the following method to assign the size to the various regions when $n > 1$. After one of the $n$ failure regions is assigned $q\%$ of $\theta$, the remaining $n-1$ failure regions share the $(100-q)\%$ as follows:

1. Randomly generate $n-1$ numbers $X_1, X_2, ..., X_{n-1}$ within the range $(0, 1)$.
2. Define $\theta_i$ for each of the $n-1$ failure regions to be
   $\frac{X_i \times \theta}{\sum_{i=1}^{n-1} X_i}$

The experiment was conducted using $q = 0, 50$ and 80. Obviously, a predominant failure region exists when $q = 50$ or 80. The results of the experiment are shown in Figure 7. Also included is the case where all failure regions are of the same size. For ease of discussion, “equal distribution” represents the case where all regions are of equal size, “uniform distribution”, “50% predominant distribution” and “80% predominant distribution” represent the cases where $q$ is 0, 50 and 80, respectively.

Figure 5. The relationship between the ART F-ratio and the number ($n$) of equally sized 2D square failure regions, where $n$ was varied from 1 to 100

Figure 6. The relationship between the ART F-ratio and the number ($n$) of equally sized 3D cuboid failure regions, where $n$ was varied from 1 to 100

Figure 7. The relationship between the ART F-ratio and the number ($n$) of failure regions for various region size distributions, where $n$ was varied from 1 to 100
As shown in Figure 7, the favourable distributions for ART are in the following order: 80% predominant distribution as the most favourable, 50% predominant distribution, uniform distribution, and equal distribution as the least favourable. Interestingly, the ART F-ratio is very steady with respect to \( n \) for the 80% (around 0.7) and 50% (around 0.85) predominant distributions. Furthermore, when \( n \) is small, the ART F-ratio under the uniform distribution is slightly smaller than that under the equal distribution, but as \( n \) increases, the two become similar. This is understandable because the larger \( n \) for the uniform distribution, the smaller the size differences among failure regions, and hence, the uniform distribution becomes more similar to the equal distribution. The results of Experiment 4 indicate that the existence of a predominant failure region apparently has a more significant impact on the ART F-ratio than does the number of failure regions.

4. Discussions and conclusion

Chan et al. [4] have observed that failure-causing inputs form different patterns: block, strip and point. The intuition of ART is to evenly spread the randomly selected test cases so it can quickly detect the first failure when the failure pattern is of block or strip type. All studies (including both simulations and experiments with real faulty programs) conducted so far support this intuition. However, the concepts of block, strip and point types were only described in general terms and this coarse classification is inadequate to precisely categorize the favourable conditions for ART (that is, conditions under which the ART F-ratio is small).

In this paper, we conducted a series of experiments to investigate into the fundamental factors influencing the fault-detection effectiveness of ART.

Our results reported in this paper reveal the nature of the different failure patterns in more precise terms in the context of ART. When there is only one failure region, the more compact it is, the more the pattern is like the block type (even when \( \theta \) is very small, it can still be of a block type as demonstrated by Experiment 1). When there is more than one failure region, the more the regions are, the more the pattern is like a point type unless there exists a predominant region.

Based on the results of Experiments 1 to 4, we conclude that the ART F-ratio depends on (1) the failure rate, (2) the number of failure regions, (3) the size of the predominant failure region, if any, and (4) the compactness of the failure region. Quantification of the first three factors is straightforward, but not the last factor. There are various metrics which can be used to measure the compactness of failure regions. It is worthwhile to investigate which metrics are more appropriate for the study of ART. So far, we have found a metric appropriate at least for the shapes that we have studied.

We have used a systematic approach to investigate into the relationship between the ART F-ratio and distributions of failure-causing inputs. We have identified four favourable conditions under which it is more cost-effective to apply ART than RT. They are: (1) when the failure rate is small, (2) when the failure region is compact, (3) when the number of failure regions is small, and (4) when a predominant region exists among all the failure regions.

It will be of great interest to know how likely these favourable conditions occur in real-life applications. Intuitively, the failure rate, number of faults and number of failure regions decrease when the software goes through a serious quality assurance process and, hence, the gain of using ART instead of RT will become greater and greater.

Acknowledgements

This research project is supported in part by an Australia Research Council Discovery Grant (DP0557246). We are most grateful to D. Towey, R. Merkel and C. Sun for their invaluable comments.

Appendix

Theorem 1

Let \( R_1 \) and \( R_2 \) be two rectangles in the 2D space and the ratios of their widths to their lengths be 1: \( r_1 \) and 1: \( r_2 \), respectively, where 1 \( \leq r_1 < r_2 \). Let their perimeters be \( P_1 \) and \( P_2 \) and their areas be \( A_1 \) and \( A_2 \), respectively, where both \( A_1 \) and \( A_2 > 0 \). If \( A_1 = A_2 \), then \[ \frac{4\pi \cdot x}{P_1} < \frac{4\pi \cdot x}{P_2} \]

\[ \Rightarrow \frac{(r_1 - r_2)(1 + \frac{1}{r_1})}{r_1} > 0 \Rightarrow r_1 + r_2 + \frac{1}{r_1} > 0 \Rightarrow r_1 + 2 + \frac{1}{r_1} > r_1 + 2 + \frac{1}{r_1} \]

\[ \Rightarrow \frac{1}{r_1} (r_1 + 1)^2 \geq \frac{1}{r_2} (r_2 + 1)^2 \Rightarrow \frac{A_1}{r_1} (r_1 + 1)^2 > \frac{A_2}{r_2} (r_2 + 1)^2 \]

(because \( A_2 = A_1 \))

\[ \Rightarrow [2 \sqrt{\frac{1}{r_1} (r_1 + 1)^2}]^2 > [2 \sqrt{\frac{1}{r_2} (r_2 + 1)^2}]^2 \Rightarrow P_2^2 > P_1^2 \Rightarrow \frac{4\pi \cdot x}{P_2} > \frac{4\pi \cdot x}{P_1} \]

Theorem 2

Let \( C_1 \) and \( C_2 \) be two cuboids in the 3D space and the ratios of their edge lengths be 1: \( r_1 \): \( r_2 \) and 1: \( r_1 \): \( r_2 \), respectively, where 1 \( \leq r_1 < r_2 \). Let their surface areas be \( P_1 \) and \( P_2 \) and their volumes be \( V_1 \) and \( V_2 \), respectively, where both \( V_1 \) and \( V_2 > 0 \). If \( V_1 = V_2 \), then \[ \frac{6A}{\sqrt{r_1} + \sqrt{r_2}} \leq \frac{6A}{\sqrt{r_1} + \sqrt{r_2}} \]

\[ \Rightarrow \frac{6A \sqrt{r_1}}{r_1} \leq \frac{6A \sqrt{r_2}}{r_2} \]

\[ \Rightarrow \frac{A_1}{r_1} \geq \frac{A_2}{r_2} \]

\[ \Rightarrow \frac{1}{r_1^{1/3}} \geq \frac{1}{r_2^{1/3}} \]

\[ \Rightarrow \frac{1}{r_1^{1/3}} > 0 \text{ and } \frac{1}{r_2^{1/3}} > 0 \]
\[ 2(r^{1/3} - r^{1/3})(r^{1/3} - 1) > 0 \Rightarrow 2(r^{1/3} - 1)(r^{1/3} - 1) > 0 \]
\[ \Rightarrow (r^{1/3} + r^{1/3})(r^{1/3} - 1) > 0 \]
\[ \Rightarrow (r^{1/3} + r^{1/3})(r^{1/3} - 1) > 0 \]
\[ \Rightarrow r^{2/3} + 2r^{1/3} - r^{1/3} > 0 \]
\[ \Rightarrow r^{2/3} + 2r^{1/3} > r^{1/3} \]
\[ \Rightarrow \frac{2}{r^{1/3}}(r^{1/3} + 2r^{1/3}) > 2(r^{1/3}) \]
\[ \Rightarrow P_2 > P_1 \Rightarrow P_2^{3/2} > P_1^{3/2} \]
\[ \Rightarrow 6.4\sqrt{P_2} > 6.4\sqrt{P_1} \]