On Test Case Distributions of Adaptive Random Testing

Tsong Yueh Chen  Fei-Ching Kuo†  Huai Liu
Faculty of ICT, Swinburne University of Technology, Australia
E-mail: {tchen, dkuo, hliu}@ict.swin.edu.au

Abstract

Adaptive Random Testing (ART) has recently been proposed as an approach to enhancing the fault-detection effectiveness of Random Testing (RT). The basic principle of ART is to enforce randomly selected test cases as evenly spread over the input domain as possible. Many ART methods have been proposed to evenly spread test cases in different ways, but no comparison has been made among these methods in terms of their test case distributions. In this paper, we conduct a comprehensive investigation on test case distributions of various ART methods. Our work shows many interesting aspects related to ART’s performance and its test case distribution. Furthermore, it points out a new research direction on enhancing ART.

1. Introduction

Random Testing (RT), a fundamental software testing technique, simply generates test cases in a random manner from the set of all possible inputs, namely the input domain [10, 14]. RT has been successfully applied in industry to detect software failures [15, 16, 17].

It has been observed that for most programs, the failure-causing inputs (program inputs that can reveal failures) are clustered together [1, 2, 9]. Chen et al. [8] studied how to improve the fault-detection effectiveness of RT under such a situation. They proposed a novel approach, namely Adaptive Random Testing (ART), where test cases are not only randomly selected from the input domain, but also enforced as evenly spread over the input domain as possible. Since then, many ART methods have been proposed, such as Fixed-Sized-Candidate-Set ART (FSCS-ART) [8], Restricted RT (RRT) [4], ART through dynamic partitioning [5] and Lattice-based ART [12]. Different ART methods distribute their test cases in different ways. All previous studies on ART methods [4, 5, 8, 12] were focused on the performance improvement that ART has over pure RT. No work has been conducted to compare ART methods with respect to their test case distributions, not to say the study on the relationship between the test case distributions and the performance of these methods.

In this paper, we study four ART methods, FSCS-ART [8], RRT [4], and two versions of ART through dynamic partitioning, namely, “by bisection” (BPRT) and “by random partitioning” (RPRT) [5]. We measure their fault-detection effectiveness, and examine their test case distributions using various metrics.

2. The effectiveness of various ART methods

For ease of discussion, we introduce notations and concepts commonly used in this paper as follows:

- $D$ denotes the input domain of $N$ dimension.
- $dD$ denotes $d$-dimension, where $d = 1, 2, \cdots, N$.
- $E$ denotes the set of already executed test cases.
- $|D|$ and $|E|$ denote the size of $D$ and $E$, respectively.
- $\theta$ denotes failure rate, ratio of the number of failure-causing inputs to the number of all possible inputs.

There are different notions of implementing ART, and different notions give rise to various ART methods. For the detailed algorithms of the ART methods, refer to [11]. The performance of ART methods is usually evaluated by F-measure, the expected number of test cases to detect the first failure. A testing method is considered more effective if it has a smaller F-measure.

As shown in [7], ART performs best when failure-causing inputs are well clustered into one single compact block (results are given in Experiment 1 of [7]). We followed the same experimental setting to study how various ART methods perform. It was expected that the data collected in this section could help us better understand the relationship between ART performance and even spreading of test cases. The experimental results are summarized in Figure 1, where x-axis denotes $\theta$, and y-axis denotes ART F-ratio, which is defined as the ratio of F-measure of ART (denoted by $F_{\text{ART}}$) to that of RT (denoted by $F_{\text{RT}}$). F-ratio measures the improvement of ART over RT. From these results, the following observations can be made.

- Almost all the experimental data show that these ART methods have larger F-measures in higher dimensional spaces for the same $\theta$.

---

*This research project is supported by an Australian Research Council Discovery Grant (DP0557246).
†Corresponding author
cases generated by these ART methods will be measured. In this paper, the distribution of test cases is introduced. Suppose $p'$ is a specific failure rate such that RRT has larger $F_{\text{ART}}$ than FSCS-ART when $\theta > v$; and RRT has smaller $F_{\text{ART}}$ than FSCS-ART when $\theta \leq v$.

There is a specific failure rate $v$ such that BPRT has larger $F_{\text{ART}}$ than RPRT when $\theta > v$; and RPRT has smaller $F_{\text{ART}}$ than BPRT when $\theta < v$.

Although all above methods aim at evenly spreading test cases, this study shows that their fault-detection effectiveness vary. The differences are mainly due to their ways of distributing test cases. In this paper, the distribution of test cases generated by these ART methods will be measured.

3. Measurement of test case distribution

For ease of discussion, the following notations are introduced. Suppose $p'$ and $p''$ are two elements of $E$. $\text{dist}(p', p'')$ denotes the distance between $p'$ and $p''$; and $p(p, E\{\{p\}\})$ denotes the nearest neighbour of $p$ in $E$. Without loss of generality, the range of values for each dimension of $D$ is set to $[0, 1)$, or simply $D = [0, 1)^N$.

In this study, three metrics were used to measure the test case distributions (the distribution of $E$ in $D$). The following outlines the definitions of these three metrics.

- **Discrepancy**

$$M_{\text{Discrepancy}} = \max_{i=1...m} \left| \frac{E_i}{|E|} - \frac{|D_i|}{|D|} \right|$$

where $D_1, D_2, ..., D_m$ denote $m$ randomly defined subsets of $D$, with their corresponding sets of test cases being denoted by $E_1, E_2, ..., E_m$, which are subsets of $E$. Note that $m$ is set to 1000 in this paper.

Intuitively, $M_{\text{Discrepancy}}$ indicates whether regions have an equal density of points. $E$ is considered reasonably equidistributed if $M_{\text{Discrepancy}}$ is close to 0.

- **Dispersion**

$$M_{\text{Dispersion}} = \max_{i=1...|E|} \text{dist}(e_i, p(e_i, E\{\{e_i\}\}))$$

where $e_i \in E$.

Intuitively, $M_{\text{Dispersion}}$ indicates whether any point in $E$ is surrounded by a very large empty spherical region. A small $M_{\text{Dispersion}}$ indicates that $E$ is reasonably equidistributed.

- **The ratio of the number of test cases in the edge of the input domain ($E_{\text{edge}}$) to the number of test cases in the central region of the input domain ($E_{\text{centre}}$)**

$$M_{\text{EdgeCentre}} = \frac{|E_{\text{edge}}|}{|E_{\text{centre}}|}$$

where $E_{\text{edge}}$ and $E_{\text{centre}}$ denote two disjoint subsets of $E$ locating in $D_{\text{edge}}$ and $D_{\text{centre}}$, respectively; $E = E_{\text{edge}} \cup E_{\text{centre}}$. Note that $|D_{\text{centre}}| = \frac{|D|}{2}$ and $D_{\text{edge}} = D \setminus D_{\text{centre}}$. Therefore, $D_{\text{centre}} = \left[ \frac{1}{2} - \frac{N}{2N+1}, \frac{1}{2} + \frac{N}{2N+1} \right]$. Clearly, in order for $M_{\text{Discrepancy}}$ to be small, the $M_{\text{EdgeCentre}}$ should be close to 1; otherwise, different parts of $D$ have different densities of points.

Discrepancy and dispersion are two commonly used metrics for measuring sample point equidistribution. More details of these two metrics can be found in [3].

The above three metrics were used to measure the test case distribution of a testing method from various perspectives. The space where a method generated points (test cases) was set to either 1D, 2D, 3D, or 4D. $|E|$ was set as from 100 to 10000. A sufficient amount of data were collected in order to get a reliable mean value of a metric within 95% confidence level and ±5% accuracy range.

It is interesting to find out how test cases of pure RT are distributed with respect to these metrics. Like all previous studies of ART, it was assumed that RT has a uniform distribution of test cases, which means that all test cases have an equal chance of being selected. Note that “uniform distribution” does not imply even spreading of test cases.

4. Analysis of test case distributions

The ranges of $M_{\text{EdgeCentre}}$ for all testing methods are summarized in Figure 2, with the following observations.

- When $N = 1$, $M_{\text{EdgeCentre}}$ for all ART methods under study is close to 1.

- When $N > 1$, FSCS-ART and RRT tend to generate more test cases in $D_{\text{edge}}$ than in $D_{\text{centre}}$ (or simply, FSCS-ART and RRT have edge bias). Moreover, the edge bias is stronger with RRT than with FSCS-ART.
When \( N > 1 \), RPRT allocates more test cases in \( D_{centre} \) than in \( D_{edge} \) (or simply, RPRT has centre bias). But the centre bias of RPRT is much less significant than the edge bias of FSCS-ART or RT.

- The edge bias (for FSCS-ART and RRT) and centre bias (for RPRT) increase as \( N \) increases.
- RT and BPRT have neither edge bias nor centre bias.

The ranges of \( M_{\text{Dispersion}} \) for all testing methods are summarized in Figure 3, with the following observations.

![Figure 2. Range of \( M_{\text{Dispersion}} \) for each testing method and dimension where \( |E| \leq 10000 \)](image)

- When \( N > 1 \), all ART methods have smaller \( M_{\text{Dispersion}} \) values than RT.
- For \( N > 1 \), RRT normally has the smallest \( M_{\text{Dispersion}} \) values, followed in ascending order by FCS-ART, BPRT, RPRT, and RT.

In Table 1, the testing methods are ranked according to their test case distribution metrics. For the same metric, the method which most satisfies the definition is ranked 1, and the one which least satisfies the definition is ranked 5. When two methods satisfy the definition to more or less the same degree, they are given the same ranking.

![Table 1. Testing methods ranked according to test case distribution metrics](image)

For those testing methods studied, in the 1D case, it has been observed that RT has the best performance, followed by FSCS-ART, BPRT, RPRT and then RT. However, when looking at the 2D, 3D and 4D cases, the same performance ordering is observed for small \( \theta \), but almost the reverse ordering for large \( \theta \). It has been shown in [11] that there exist some hidden factors (unrelated to how evenly a method...
spreads test cases) which have a strong impact on the performance of ART. Only when \( \theta \) is small enough, will the performance of ART strongly depend on how evenly spread its test cases are. In order to fairly analyze the relationship between the test case distribution and the performance of ART methods, without being influenced by external factors, the rest of the discussion will be carried out on small failure rates.

Table 1 shows that in terms of \( M_{\text{Dispersion}} \) metric, RRT has the most even spreading of test cases, followed by FSCS-ART, BPRT and RPRT. In other words, the ranking according to the \( M_{\text{Dispersion}} \) metric is consistent with the ranking according to F-measures (data shown in Figure 1). It should be pointed out that even though \( M_{\text{Dispersion}} \) and \( M_{\text{Dispersion}} \) are two commonly used metrics for measuring sample point equidistribution, in this study, \( M_{\text{Dispersion}} \) appears to be more appropriate than \( M_{\text{Dispersion}} \).

Interestingly, among all ART methods under study, the one with the largest \( M_{\text{Edge:Centre}} \) (that is, RRT) has the smallest \( M_{\text{Dispersion}} \), while the one with the smallest \( M_{\text{Edge:Centre}} \) (that is, RPRT) has the largest \( M_{\text{Dispersion}} \). As discussed before, \( M_{\text{Dispersion}} \) could best reflect the ordering of testing methods with respect to their performance. We notice that pushing test cases away (so that \( M_{\text{Edge:Centre}} > 1 \)) is not a bad approach to evenly spreading test cases, even though it may not be the best approach to achieving a real even spreading of test cases.

5. Conclusion

Previous studies [4, 5, 8] showed that even spreading of test cases makes ART outperform RT. The concept of even spreading of test cases is simple but vague. In this paper, several metrics were used to measure the test case distribution of ART as well as RT. The relevance and appropriateness of these metrics were also investigated. To our best knowledge, this is the first work on analyzing the relationship between test case distributions and performance of an ART method. Recently, there were some works on alleviating the edge bias of FSCS-ART [6, 13]. We shall continue the line of this research with additional knowledge gained from this study to enhance the existing ART methods.

References