

Statistical Analysis and Enhancement of Random Testing Methods also under Constrained Resources

Johannes Mayer, Christoph Schneckenburger
Ulm University

Dept. of Applied Information Processing
89069 Ulm, Germany

Email: {johannes.mayer,christoph.schneckenburger}@uni-ulm.de

Abstract—Adaptive Random Testing (ART) denotes a family of random testing methods that are designed to be more effective than Random Testing (RT). Mostly, these methods have been investigated using the mean F-measure, which denotes the random number of test cases necessary to detect the first failure. The two most important ART methods, namely Distance-Based ART (D-ART) and Restricted Random Testing (RRT), perform worse for higher failure rates than for lower failure rates. Furthermore, all previous publications on ART analyzed these methods for testing with unlimited resources. The present paper investigates, why D-ART and RRT behave better for lower failure rates. Therefore, the F-measure distribution and the spatial distribution of single test cases are analyzed. Thereby, shortcomings of D-ART and RRT are revealed. Improved ART methods are presented based on our findings. Furthermore, the usefulness of the F-measure distribution of testing with unlimited resources for resource-constrained testing is explained. Finally, the ART methods are compared to RT for both cases, i.e. with and without resource limitations.

Keywords: F-measure, Distribution, Adaptive Random Testing, Random Testing, Resource-constrained testing

I. INTRODUCTION

Software testing can be defined as the process of executing a program in order to reveal bugs [1]. Besides other modern definitions, this “old” definition is still valid in most cases. Since it is usually quite time consuming to produce a sufficient number of test cases, Random Testing (RT) [2], [3], [4], [5], [6], i.e. the random generation of test cases, has gained much importance. Random Testing has successfully been applied to test database systems [7], the robustness of Windows NT applications [8], and Java Just-In-Time (JIT) compiler [9]. Using Random Testing, it is quite simple to automate testing—given that an oracle is accessible. Random Testing, however, makes no assumption on the program under test [1]. Therefore, Chen et al. [10] introduced Adaptive Random Testing (ART). Since failure causing inputs mostly appear clustered within the input domain—Chan et al. [11] roughly categorized common failure patterns into block, strip, and point pattern—, ART methods implement the (informal) notion of wide-spread test cases in order to achieve a better performance than RT. This can be accomplished for example through distance computations resp. restriction, implemented by Distance-Based Adaptive Random Testing (D-ART) [10] resp. Restricted Random Testing (RRT) [12], [13], [14].

In order to compare random testing methods, Chen et al. introduced the *F-measure* which denotes the (random) number of test cases necessary to exhibit the first failure. This measure is very intuitive and natural, since testing is often stopped when a failure has been detected. Most ART methods have been compared using the mean F-measure. Thereby, it turned out among others that D-ART and RRT perform better for lower failure rates than for higher failure rates [15]. However, the validity of an analysis only based on the mean F-measure is questionable. Chen et al. [16], [17] did a first investigation of the distribution of the F-measure to compare D-ART, RRT, and RT. They have shown that the F-measure is geometrically distributed for RT. Since theoretical statements about the F-measure distribution of both D-ART resp. RRT seem to be impossible, they presented some empirical study based on simulations for D-ART, RRT, and RT and reasoned about the validity of comparisons based on the mean F-measure.

The present paper first investigates the important problem of worse performance of D-ART and RRT for higher failure rates. Therefore, the F-measure distribution is analyzed in more detail than in [16], [17]. Furthermore, the spatial distribution of test cases is also considered. Based on our investigation, improved ART methods are presented. The present paper also analyzes resource-constrained testing—the usual case. This has not been considered in all publications on ART. It will be shown, how the F-measure distribution determined for testing without resource limitations can be used to acquire information about resource constrained testing. Thereby, ART and RT can be compared for the cases with and without resource limitations.

In the following section, some necessary preliminaries concerning the investigated ART methods, the notation, and testing effectiveness measures are introduced. Section III investigates the distribution of the F-measure through a study, discusses the results, and gives explanations for shortcomings of D-ART resp. RRT supported by further results. Based on these shortcomings, Section IV introduces a more effective modification of the common D-ART and RRT methods. The usefulness of the distribution of the F-measure for resource-constrained testing is explained in Section V, followed by a conclusion and perspectives for future research (Section VI).

II. PRELIMINARIES

A. Investigated ART Methods

The first ART algorithm was Distance-Based ART (D-ART) [10], [18] with fixed sized candidate set. This method chooses the first test case purely randomly. Thereafter, a set of test case candidates—each chosen purely randomly—of size k is computed in each iteration. That element from the candidate set with the greatest minimal Euclidean distance to the already executed test cases is then chosen as the next test case and executed. In the following iteration, a new candidate set is chosen and the procedure is repeated until a failure is detected (or the resources for testing are exhausted). The size $k = 10$ of the candidate set has been recommended in the above cited publications.

Another ART method is based on the notion of restriction, namely Restricted Random Testing (RRT) [12], [13], [14]. The first test case is again chosen purely randomly. In each iteration a disc with exclusion radius $r := \sqrt{A \cdot R / (\pi \cdot n)}$ is located around each previously executed test case that did not exhibit a failure. A denotes the area of the input domain, R denotes the coverage ratio (i. e. the area of the exclusion zone related to the area of the input domain), and n is the number of previously executed test cases not causing a failure. Each randomly generated test case candidate that falls into one of these circular exclusion zones is rejected. The first candidate, that does not fall into any exclusion zone, is chosen as the next test case and—if executing this test case yields no failure—the procedure is repeated. A coverage ratio $R = 1.5$ is given as a recommendation.

B. Notation

Those inputs from the input domain which exhibit a failure are called *failure-causing inputs*. The *failure rate* θ is then defined as the percentage of failure-causing inputs within the input domain. For an input domain with area A and a failure rate θ the area of the failure pattern is θA .

C. The F-Measure Distribution of Random Testing

For Random Testing with uniform operational profile and replacement, the F-measure is geometrically distributed [16], [17]. Therefore, the probability that exactly k test cases are necessary to detect the first failure is $\mathbb{P}(F = k) = \theta(1 - \theta)^{k-1}$ for $k \in \mathbb{N}^+$ and the theoretical mean F-measure is equal to $1/\theta$ for RT. For example, the theoretical mean F-measure of Random Testing with replacement is 100 if the failure rate is $\theta = 0.01$. In order to compare the F-measure for different failure rates, we use the *relative mean F-measure*, defined as the mean F-measure related to the theoretical mean F-measure of Random Testing.

D. Other Testing Effectiveness Measures

Further common testing effectiveness measures are the P-measure and the F-measure. The *P-measure* is defined as the probability that at least one failure is detected with a specified test set whereas the *E-measure* denotes the expected number of failures detected by the test set. Both seem to be less

appropriate to compare random testing methods, since the test set of RT methods is not fixed but steadily growing. Furthermore, the F-measure is more intuitive and well-known in the field of ART methods.

III. EMPIRICAL STUDY

A. Simulation Design

Chen et al. [16], [17] chose a sample size of 2,000 for their study, i. e. the algorithms were run with 2,000 randomly chosen failure patterns. They presented their results for the failure rate $\theta = 0.002$, since they argued that varying the failure rate has negligible effect on the distribution (apart from obvious scaling effects). Furthermore they chose relatively big histogram classes of 285 test cases.

In order to refine the study described in [16], [17], a single histogram class was used for each possible value of the F-measure and a sample size of 45,000,000 was chosen. In our study, the failure rate θ was chosen 0.01 to keep the run-time as small as possible. The behavior for smaller failure rates (e.g. $\theta = 0.002$ as Chen et al. [16], [17] used) will be discussed afterwards.

Moreover, using the Central Limit Theorem [19], confidence intervals on confidence level 99% can be established for each value of the empirical density function. The value of each histogram class is therefore modeled as a random variable being the mean of i. i. d. 0/1-valued random variables. Thus, the empirical density functions can be plotted with error bars, which are necessary to estimate the quality of the empirical result. Such error bars are missing in [16], [17].

The failure patterns were generated as done by Chen et al. [16], [17]. That means, that the block failure pattern was generated by a random square of size θA totally within the input domain. For the strip pattern, two adjacent sides and two points on these sides were chosen randomly. Then the strip was constructed centered on the line connecting these points, and its width was computed such that the strip had the desired area θA . Points near the corners were rejected to avoid overly wide strips. For the point pattern, 10 non-overlapping discs with equal radius lying totally within the input domain were randomly generated to achieve the total area θA .

B. Results and Discussion

Figure 1 shows the empirical density function of the F-measure for Distance-Based Adaptive Random Testing (D-ART) as well as for Restricted Random Testing (RRT) for the block failure pattern. For comparison purposes, the theoretical density function of Random Testing is depicted as well—which is geometric as Chen et al. [16], [17] have proven.

Since the first test case of both D-ART and RRT is chosen purely randomly, it is not surprising, that its probability of detecting a failure is the same as that of Random Testing. The second test case performs significantly worse than the first test case and even than the second test case of Random Testing. After the second test case, the probability of detecting a failure is more or less moderately increasing up to about the 30th (D-ART) resp. the 40th (RRT) test case. From this local maximum

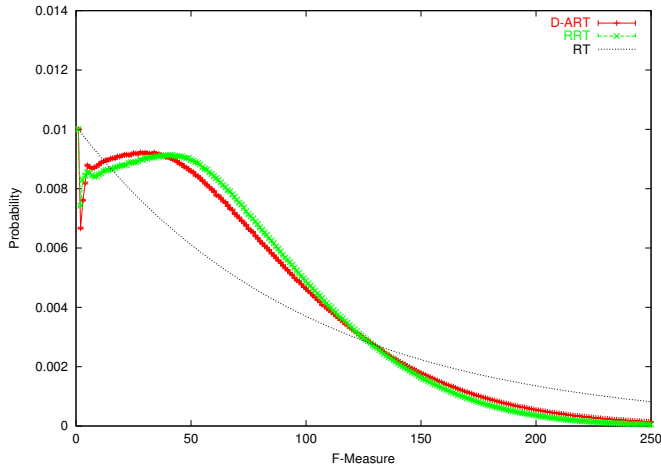


Fig. 1. Empirical density function of the F-measure for D-ART and RRT in contrast to the theoretical density function of RT for the block failure pattern and failure rate 0.01

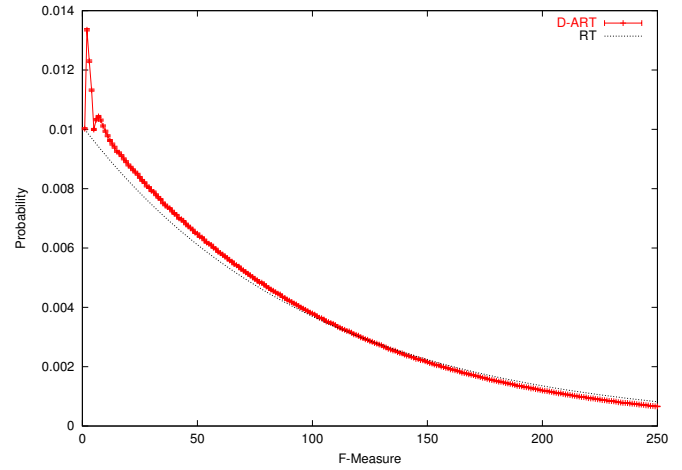


Fig. 2. Empirical density function of the F-measure for D-ART in contrast to the theoretical density function of RT for the strip failure pattern and failure rate 0.01

on, the empirical density function decreases, resembling the density function of a geometric distribution.

For the block failure pattern, there are at least two facts worth mentioning. On the one hand, the failure finding effectiveness of both ART methods is worse than that of pure Random Testing for about the first 15–20 test cases. On the other hand, there is a (local) maximum up to which the probability of detecting a failure is increasing and after which this probability is decreasing. This maximum is about the 30th test case for D-ART and about the 40th for RRT. Before we identify the reasons for this behavior we will have a look at the other failure patterns.

Figure 2 shows the empirical density function of the F-measure for the strip failure pattern. As before, the theoretical density function of Random Testing is depicted as well. Since the empirical density functions of D-ART and RRT are very similar, Figure 2 solely depicts the density function of D-ART. For the strip failure pattern the probability of detecting a failure increases abruptly for the second test case and decreases immediately thereafter. From about the 10th test case on, the empirical density function resembles the density of a geometric distribution.

For the point failure pattern, the empirical density function of D-ART, as well as the theoretical density function of Random Testing, are illustrated in Figure 3. Here the differences to Random Testing are marginal, except from the second test case, for which the probability of detecting a failure is significantly worse, and some further test cases. But despite this similarity to the geometric distribution, there is one fact mentionable. For the first about 40–50 test cases D-ART performs worse than RT and after these test cases up to about the 200th test case D-ART performs better. The empirical density function for the point pattern is quite independent of the chosen method (D-ART or RRT) and even—as further studies have shown—of the number of points (e. g. 10 or 50).

In order to determine the reason for the especially low

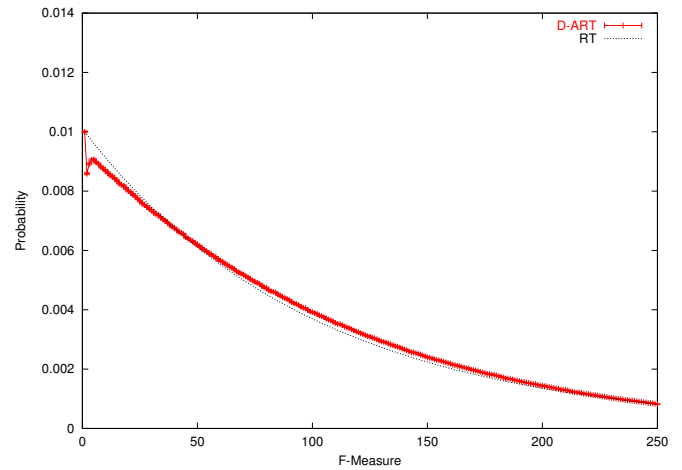


Fig. 3. Empirical density function of the F-measure for D-ART in contrast to the theoretical density function of RT for the point failure pattern and failure rate 0.01

resp. high effectiveness of the first test cases—according to the respective failure pattern—, the position of the i th generated test case was analyzed. For this purpose the first 100 test cases of D-ART and RRT were generated 10,000,000 times as if a correct program was tested. The results are, however, also applicable to faulty programs, since at least the first test cases seldomly detect a failure, esp. in case of lower failure rates. Figure 4 depicts the results as a two-dimensional (spatial) density function where brighter pixels represent higher probability of occurrence. Since there were no significant differences between D-ART and RRT, Figure 4 only depicts the spatial distribution of the F th test case of D-ART.

Since the first test case is chosen purely randomly, the figure shows an equal distribution of the first test case ($F = 1$). However, the second test case ($F = 2$) obviously prefers the corners of the input domain. This seems to be the reason why the second test case (of D-ART and RRT) is not so effective

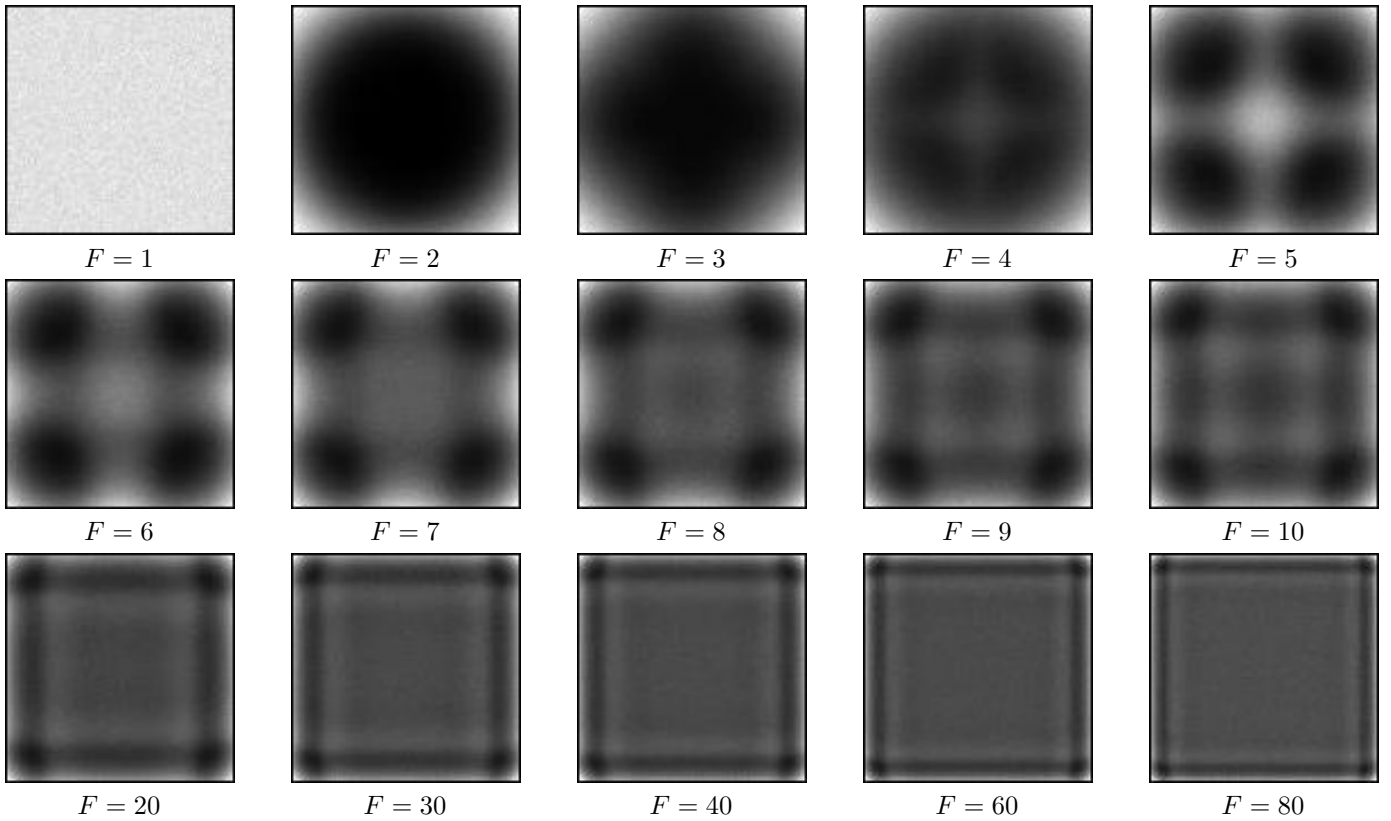


Fig. 4. Spatial distribution of the F th test case of D-ART

for block and point failure patterns. Since the strip failure pattern always connects two sides of the input domain, it always extends to the boundary and partly even to the corners of the input domain. Thus, it is not surprising, that D-ART and RRT are very effective with the second test case. The spatial distribution of the third and fourth test case is roughly the same as that of the second test case. From the fifth test case on, the center of the input domain is chosen significantly more frequently. The distribution of the fifth to the tenth test case show some lattice-like preference and after these test cases, the preferences to some regions of the input domain are diminishing. But still for the 80th test case some kind of frames within the input domain are noticeable. Test cases at the boundary of the input domain are preferred. Adjacent to this region is a darker frame, which has very low probability of occurrence. Test cases inside this dark frame are distributed quite equally.

Due to the fact that D-ART and RRT try to maximize the distance between a test case and all previously executed test cases, the preference of the corners and the boundary for the first test cases can be explained. Furthermore, the “dark frame” (e. g. for $F = 80$) can be explained by the fact that there are already many test cases at the boundary of the input domain and therefore test case candidates in the “darker frame” will often be rejected resp. be replaced by another one, since they are too close to the “boundary test cases”.

The spatial distribution of the test cases clearly illustrates

why the first test cases behave especially well or worse, according to the kind of failure pattern. Since even the “later” test cases (e. g. $F = 80$) show some kind of preference to some regions of the input domain—in particular the boundary—the method cannot be optimal for failure patterns with arbitrary position within the input domain.

The second notable fact is the local maximum, up to which the density function increases and after which it decreases, resembling the density of a geometric distribution. Further investigations have shown that this maximum can be moved horizontally by varying the number of test case candidates (D-ART) resp. the coverage ratio (RRT). However, a reduction in candidate set size resp. in the coverage ratio weakens the influence of the first test cases, such that the performance of the method enhances for the first test cases.

The question arises, whether the distribution of the F-measure is basically the same for smaller failure rates θ , e. g. for $\theta = 0.002$ as Chen et al. have chosen in [16], [17]. For this purpose, we did further investigations for other failure rates. The number of the test cases after which Adaptive Random Testing is more effective than Random Testing does not vary much in the failure rate (apart from obvious scaling factors). Since for smaller failure rates the first test cases become less important, the fact that the F-measure of both D-ART and RRT decreases for decreasing failure rates seems to be explained. However, the position of the local maximum varies in the failure rate. That means, that the “frontier test case”, the test

case after which the empirical density function decreases, is about the same for all failure rates θ up to a scaling factor.

IV. CONTINUOUS ADAPTIVE RANDOM TESTING

A. The Method

Our study has clearly shown, that one major shortcoming of distance-based resp. restriction-based ART versus RT is the low effectivity of the first test cases (especially for the block failure pattern) based on the spatial distribution of individual test cases depicted in Figure 4. Since this behavior seems to affect the performance of the whole method, we introduce a slight modification of the distance computation in order to achieve high effectivity even for the first test cases. This modification is based on the idea of widespread test cases, which is the fundamental concept of Adaptive Random Testing [10]. The results depicted in Figure 4 have shown that not all test cases are evenly spread. Test cases close to the corners and boundary are more probable at the beginning. The reason for this is that there are less neighbors close to the boundary and corners. Therefore, less test cases candidates are rejected or replaced by other candidates at the boundary. Consequently, we try to modify the distance computation in D-ART and RRT such that all points within the input domain have virtually the same number of neighbors in order to achieve a better even spread of the test cases.

Instead of computing distances just within the input domain, the input domain is regarded as virtually continuous, such that the left border is directly adjacent to the right border, the upper border to the lower, and also the upper-right corner to the lower-left corner and so on. Then, the distances can also be computed “around” the input domain. Based on the minimum of all possible distances D-ART resp. RRT can be applied as introduced in [10], [13].

An easier way to imagine this procedure is depicted in Figure 5. We can imagine that the input domain is surrounded by copies (depicted gray) of itself. Figure 5 only depicts the relevant copies in this case, but we can imagine that the whole plane is covered by such copies. The distance (e. g. between $P1$ and $P2$) is then computed as the minimum distance between $P1$ and $P2$, where $P1$ resp. $P2$ can also be replaced by the respective copies of $P1$ resp. $P2$ —it makes only sense to replace at most one point by one of its copies. In Figure 5, the (minimum) distance between $P1$ and $P2$ is achieved between $P1$ and the copy of $P2$ in the left-upper gray copied domain. Therefore, $d2$ is the distance between $P1$ and $P2$ in this case.

We now describe the modified way of distance computation more formally. In an d -dimensional input domain, the (Euclidean) distance between $a = (a_1, \dots, a_d)$ and $b = (b_1, \dots, b_d)$ with $a, b \in \mathbb{R}^d$ can be computed as follows:

$$dist(a, b) = \sqrt{\sum_{i=1}^d (a_i - b_i)^2}$$

In contrast to this, we compute the distance within the virtually

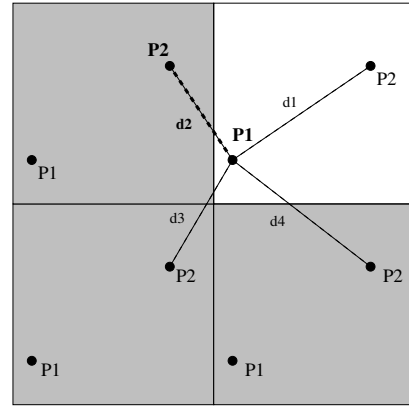


Fig. 5. Distance computation in a virtually continuous input domain

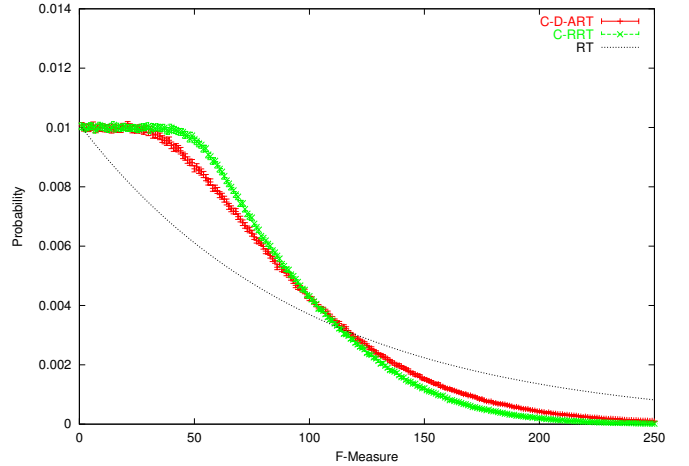


Fig. 6. Empirical density function of the F-measure of C-D-ART and C-RRT in contrast to the theoretical density function of the F-measure of RT for the block failure pattern and failure rate 0.01

continuous domain as follows:

$$dist(a, b) = \sqrt{\sum_{i=1}^d (\min(|a_i - b_i|, (w_i - |a_i - b_i|)))^2},$$

where $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ denotes the width of the (rectangular) input domain in the corresponding dimension.

The virtual input domain composed of the original domain and its copies filling the whole plane is unrestricted and does not contain corners or a boundary which might be preferred by test cases. Therefore, we call this improved method Continuous Adaptive Random Testing (C-ART). The above idea of distance computation can be applied to D-ART as well as RRT, yielding the improved methods C-D-ART and C-RRT.

B. Analysis of the F-Measure Distribution

Figure 6 depicts the empirical density functions of Continuous D-ART (C-D-ART) resp. Continuous RRT (C-RRT) for the block failure pattern and failure rate 0.01. The simulation was done as specified in Section III-A, with the difference that the sample size was chosen 5,000,000.

When comparing the empirical density functions in Fig. 1 and Fig. 6 it becomes obvious that for the block failure pattern the empirical density function of Continuous D-ART (C-D-ART) resp. Continuous RRT (C-RRT) is above both that of D-ART resp. RRT, for the first 50 test cases, and moreover that of RT. The second remarkable fact is, that the “frontier test case”, after which the probability of detecting a failure decreases, seems to be the same as the local maximum for D-ART resp. RRT (about 30 for C-D-ART and 40 for C-RRT). But whereas for the second test case D-ART and RRT become significantly less effective and moderately advance afterwards, the continuous variants (C-D-ART and C-RRT) perform relatively constant up to the “frontier test case”. Further investigations have shown, that the position of the “frontier test case” again depends on the size of the candidate set (C-D-ART) resp. the coverage ratio (C-RRT). But here the number of test candidates resp. the coverage ratio determines not only the effectivity of the first test cases but also the position of the “frontier test case” after which the density function decreases abruptly.

The spatial distribution of all test cases of C-ART look like the first pattern ($F = 1$) of Fig. 4. That means that Continuous ART has no preferences within the original input domain and therefore eliminates one major drawback of D-ART and RRT.

C. Analysis of the Mean F-Measure

Since the failure finding effectiveness seems to be better for C-ART than D-ART resp. RRT—especially for the first test cases—, the mean F-measure of C-D-ART resp. C-RRT should be lower than that of D-ART resp. RRT. Moreover, since the first test cases cannot affect differently on different failure rates, there should be no differences in the mean F-measure for various failure rates—at least for the block and the point pattern. The strip pattern behaves a little bit different since a reduction in the failure rate affects only the width of the strip and not the length. Therefore, the randomly generated strips get narrower for lower failure rates such that all common ART methods are less effective for lower failure rates. For the following simulation, the sample size was chosen 10,000. Using the Central Limit Theorem [19] again, confidence intervals can be computed for each relative empirical F-measure (on confidence level 99%). Table I shows the empirical relative mean F-measure for the block failure pattern and some failure rates.

As supposed, the mean F-measure of C-ART is strictly lower than that of D-ART resp. RRT for the block failure pattern and seems furthermore to be constant in the failure rate θ . Further studies have shown, that for the strip pattern C-D-ART resp. C-RRT perform at least not significantly worse than the common methods. For point pattern—independent of the number of points—, the Continuous ART methods outperform the common ones. Moreover the F-measure is again constant in the failure rate for the point failure pattern.

TABLE I
EMPIRICAL RELATIVE MEAN F-MEASURE FOR DIFFERENT ART METHODS
AND THE BLOCK FAILURE PATTERN

	D-ART	RRT	C-D-ART	C-RRT
$\theta = 0.01$	0.678 (± 0.013)	0.648 (± 0.012)	0.628 (± 0.012)	0.582 (± 0.011)
$\theta = 0.005$	0.662 (± 0.013)	0.632 (± 0.012)	0.620 (± 0.012)	0.586 (± 0.011)
$\theta = 0.002$	0.647 (± 0.013)	0.606 (± 0.011)	0.610 (± 0.012)	0.580 (± 0.011)
$\theta = 0.001$	0.642 (± 0.013)	0.599 (± 0.011)	0.625 (± 0.013)	0.573 (± 0.011)
$\theta = 0.0005$	0.637 (± 0.013)	0.587 (± 0.011)	0.627 (± 0.013)	0.585 (± 0.011)

V. RESOURCE-CONSTRAINED TESTING

All studies of ART methods so far have been conducted with simulated failure patterns and mutated programs. Therefore, a failure can be detected. The situation of a bug-free program has, however, been excluded. One can argue that every program contains a bug. That is true in most cases, but one fact complicates the situation: In practice, one has only a fixed budget of money and time, i. e. resources are limited. Consequently, it may be possible that a program contains bugs, but they are not found during the test. The mean F-measure of ART with unlimited resources has no meaning for resource constrained-testing. However, the distribution of the F-measure for ART with unlimited resources can be used to gain valuable information. Assume that the number of test cases which can be executed is limited by N_{max} . Let $\mathbb{P}(F = i)$ denote the probability that a failure is detected by the i th test case the first time (i. e. no failure is exhibited by the previous test cases). Then, the value of the cumulative distribution function (cdf) of the F-measure at N_{max} —i. e. the probability that a failure is exhibited even with limited resources—is $\sum_{i=1}^{N_{max}} \mathbb{P}(F = i)$. An analysis of ART methods with the F-measure distribution for unlimited resources is thus also useful for resource-constrained testing. Figures 7–9 depict the empirical cdf of the F-measure of D-ART (RRT similar) without resource limitation.

We observe that for the failure rate 0.01 the cdf of D-ART and RRT is above the cdf of RT starting at F-measure about 30 for the block failure pattern, always for the strip failure pattern, and starting at F-measure about 120 for the point failure pattern. The test case, after which the cdf of D-ART is above that of RT seems to be constant in the failure rate θ . This explains—as mentioned—the decreasing mean F-measure of D-ART resp. RRT for decreasing failure rate, since these first test cases gain less importance for lower failure rates and consequently higher F-measure.

Further investigations have shown, that the cdf of both C-D-ART and C-RRT is strictly above that of Random Testing, independent of the failure pattern and the failure rate (cf. Figure 10 for the cdf for the block failure pattern).

Summarizing, the D-ART and RRT methods are better than Random Testing in the resource-constrained case for most

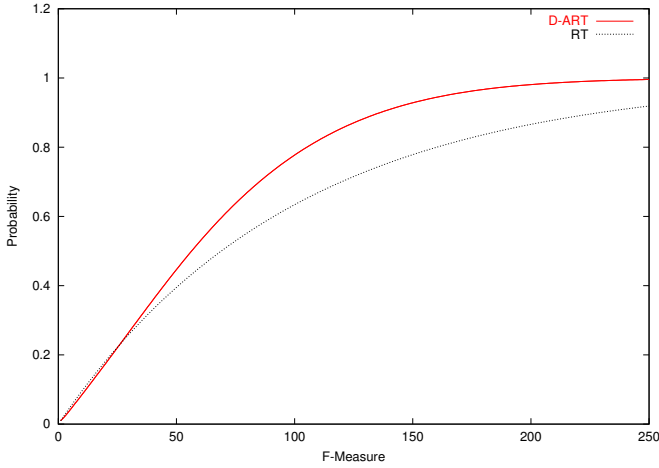


Fig. 7. Empirical cdf of the F-measure of D-ART in contrast to the theoretical cdf of the F-measure of RT for the block failure pattern and failure rate 0.01

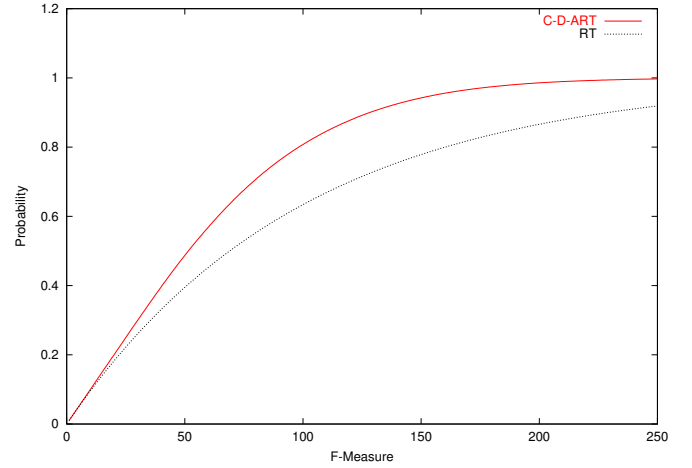


Fig. 10. Empirical cdf of the F-measure of C-D-ART in contrast to the theoretical cdf of the F-measure of RT for the block failure pattern and failure rate 0.01

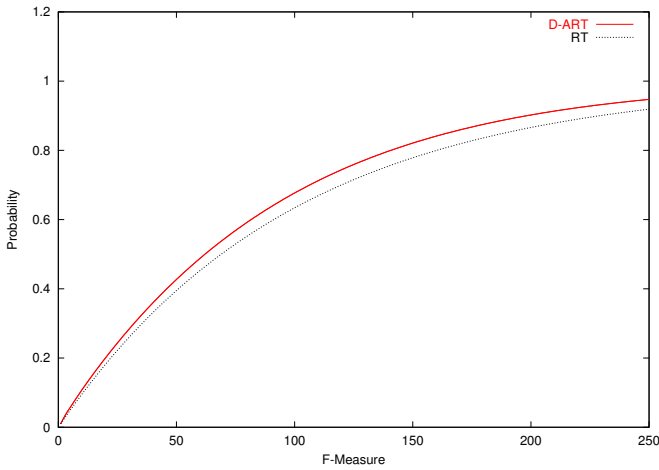


Fig. 8. Empirical cdf of the F-measure of D-ART in contrast to the theoretical cdf of the F-measure of RT for the strip failure pattern and failure rate 0.01

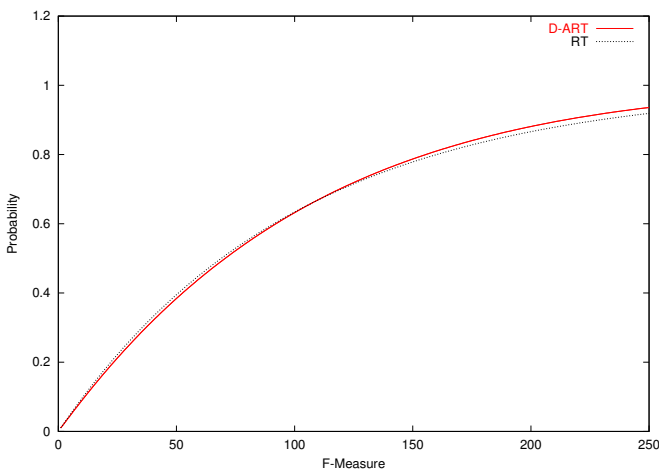


Fig. 9. Empirical cdf of the F-measure of D-ART in contrast to the theoretical cdf of the F-measure of RT for the point failure pattern and failure rate 0.01

common failure patterns, the block and the strip pattern, after only a small number of test cases. And for smaller numbers of test cases these ART methods are at least not much worse than Random Testing for the common failure patterns. Moreover the introduced Continuous ART methods are strictly better than Random Testing starting from the first test case. D-ART resp. RRT and especially C-DART resp. C-RRT should thus be preferred over RT in the resource-constrained case—as well as in the case with unlimited resources—and distributional results on the F-measure for the case without resource limitations can be used to gain useful information for the situation with limited resources.

VI. CONCLUSION AND PERSPECTIVES

In the present paper, we have investigated the distribution of the F-measure, i.e. the (random) number of test cases necessary to detect the first failure, in order to identify shortcomings of common ART methods, namely D-ART and RRT. Our study used a much larger sample size and an own histogram class for each possible value of the F-measure—in contrast to the study described in [16], [17]. Therefore, we could refine the geometric distribution hypothesis of the cited publications. For the first test cases, D-ART and RRT behave better resp. worse than RT depending on the individual failure pattern. This behavior can be observed with the density function of the F-measure.

In order to obtain the reason for this behavior, we determined the spatial distribution of the F th test case of D-ART resp. RRT. Thereby, we found out that the second test case and some further test cases are mostly close to the corners resp. the boundary. Therefore, the performance of the second and further test cases can be explained. It is a consequence of the reduced number of neighbors at the corners and at the boundary.

As a result of our investigation, we proposed improved ART algorithms that regard the input domain as virtually

continuous, where copies of the original input domain are placed around it. The distance between a test case and a candidate can then be computed in this virtual domain also between copies. Thereby, each test case has the same number of neighbors, and corners and a boundary which could be preferred are thus eliminated. It turns out that the improved ART methods are superior to D-ART and RRT regarding the mean F-measure as well as the F-measure distribution. Now, the mean F-measure is independent of the failure rate (for the most common failure patterns), and the F-measure distribution is always better than D-ART, RRT, and RT as the cdfs reveal.

Finally, we investigated the usefulness of the F-measure distribution for resource-constrained testing. Although the mean F-measure results, which are presented in most ART publications, cannot be transferred from the case of unlimited resources to resource-constrained testing, the F-measure distribution can be used to gain valuable information for resource-constrained testing, namely the probability of finding a failure with limited resources.

Preliminary investigations have shown that some other ART methods (e. g. those presented in [20], [21]) are still more effective than C-ART for the first test cases. This seems to allow some kind of combination. The first test cases could be chosen according to another strategy. Thereafter, one could proceed with D-ART resp. RRT. This should lead to an improved method. The interesting questions are, however, which methods are suitable for this combination and how to find the right time to switch between the methods, since the failure rate is not known in general.

REFERENCES

- [1] G. J. Myers, *The Art of Software Testing*. New York: Wiley, 1979.
- [2] V. D. Agrawal, "When to use random testing," *IEEE Transactions on Computers*, vol. 27, pp. 1054–1055, 1978.
- [3] J. W. Duran and S. C. Ntafos, "An evaluation of random testing," *IEEE Transactions on Software Engineering*, vol. 10, pp. 438–444, 1984.
- [4] R. Hamlet, "Random testing," in *Encyclopedia of Software Engineering*. Wiley, 1994, pp. 970–978.
- [5] P. S. Loo and W. K. Tsai, "Random testing revisited," *Information and Software Technology*, vol. 30, pp. 402–417, 1988.
- [6] P. B. Schneck, "Comment on "when to use random testing"," *IEEE Transactions on Computers*, vol. 28, pp. 580–581, 1979.
- [7] D. Slutz, "Massive stochastic testing of SQL," in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB 1998)*, 1998, pp. 618–622.
- [8] J. E. Forrester and B. P. Miller, "An empirical study of the robustness of Windows NT applications using random testing," in *Proceedings of the 4th USENIX Windows Systems Symposium*, 2000, pp. 59–68.
- [9] T. Yoshikawa, K. Shimura, and T. Ozawa, "Random program generator for Java JIT compiler test system," in *Proceedings of the 3rd International Conference on Quality Software (QSIC 2003)*. IEEE Computer Society, 2003, pp. 20–24.
- [10] T. Y. Chen, H. Leung, and I. K. Mak, "Adaptive random testing," in *Proceedings of the 9th Asian Computing Science Conference (ASIAN 2004)*, ser. Lecture Notes in Computer Science, M. J. Maher, Ed., vol. 3321. Springer, 2004, pp. 320–329.
- [11] F. T. Chan, T. Y. Chen, I. K. Mak, and Y. T. Yu, "Proportional sampling strategy: Guidelines for software testing practitioners," *Information and Software Technology*, vol. 38, pp. 775–782, 1996.
- [12] K. P. Chan, T. Y. Chen, F.-C. Kuo, and D. Towey, "A revisit of adaptive random testing by restriction," in *Proceedings of the 28th International Computer Software and Applications Conference (COMPSAC 2004)*. IEEE Computer Society, September 2004, pp. 78–85.
- [13] K. P. Chan, T. Y. Chen, and D. Towey, "Restricted random testing," in *Proceedings of the 7th European Conference on Software Quality (ECSQ 2002)*, ser. Lecture Notes in Computer Science, J. Kontio and R. Conradi, Eds., vol. 2349. Springer, 2002, pp. 321–330.
- [14] —, "Normalized restricted random testing," in *Proceedings of the 18th Ada-Europe International Conference on Reliable Software Technologies*, ser. Lecture Notes in Computer Science, vol. 2655. Springer, 2003, pp. 368–381.
- [15] T. Y. Chen, F.-C. Kuo, R. G. Merkel, and S. P. Ng, "Mirror adaptive random testing," *Information and Software Technology*, vol. 46, pp. 1001–1010, 2004.
- [16] T. Y. Chen, F.-C. Kuo, and R. Merkel, "On the statistical properties of the F-measure," in *Proceedings of the 4th International Conference on Quality Software (QSIC 2004)*. IEEE Computer Society, September 2004, pp. 146–153.
- [17] —, "On the statistical properties of testing effectiveness measures," *The Journal of Systems and Software*, 2005, (in press).
- [18] T. Y. Chen, T. H. Tse, and Y. T. Yu, "Proportional sampling strategy: A compendium and some insights," *The Journal of Systems and Software*, vol. 58, pp. 65–81, 2001.
- [19] G. Casella and R. L. Berger, *Statistical Inference*. CA, USA: Wadsworth Group, Duxbury, 2002.
- [20] J. Mayer, "Lattice-based adaptive random testing," in *Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering (ASE 2005)*. New York: ACM, 2005, pp. 333–336.
- [21] —, "Adaptive random testing by bisection with restriction," in *Proceedings of the Seventh International Conference on Formal Engineering Methods (ICFEM 2005)*, ser. Lecture Notes in Computer Science, vol. 3785. Springer-Verlag, Berlin, 2005, pp. 251–263.