# 适应性分区测试

## 摘要

动态随机测试(DRT)策略利用软件的控制理论改进了传统的随机测试策略的效率。DRT 策略的主要思想是在测试期间利用历史的测试信息动态的改变测试剖面，使得具有较高故障检测能力的分区更有可能被选到。但是 DRT 策略没有考虑每次根据测试结果调整概率的幅度应该是不同的，并且每一个分区被选择的概率容易受其它分区测试结果的影响，有可能影响 DRT 策略的测试效率。另一方面，该策略的测试效率还受到分区数目以及初始测试剖面的影响。这篇文章提出两种测试策略命名为 MAPT、RAPT 策略。MAPT 策略利用 Markov 状态转移矩阵，将分区当成状态，使得某一个分区被选择的概率只跟该分区内的测试用例的执行情况有关并且概率的调整幅度由当前分区的被选择概率决定。RAPT 策略提出了另一种奖励惩罚的机制。通过实验探究初始剖面以及分区数目对 DRT 策略的影响。数据表明当分区数目较多，初始概率分布为均等分布时，DRT、MAPT、RAPT 策略具有较高的故障检测效率，并且 MAPT、RAPT 策略的测试效率比随机测试、随机分区测试以及动态随机测试的效率高。

## 1 背景介绍

随机测试[1]以及分区测试[2,3,4]是两个非常著名的测试策略。在传统的随机测试中，按照一致或者不一致的概率分布从软件输入域中选择测试用例并执行。分区测试涉及到一簇的测试技术：状态测试、数据流测试、分枝测试、变异测试等，任何的一个输入域的子域，都需要从中挑选至少一个测试用例。

Cai[5,6]等人将随机测试与分区测试结合，提出了随机分区测试策略。该策略假设待测软件的输入域被分为m个子分区。随机分区测试策略首先根据测试剖面$\{p_1, p_2, \dots, p_m\}$选择一个分区$c_i$。然后在$c_i$中随机地选择一个测试用例执行。在整个的测试过程中测试剖面的大小不变。

在随机分区测试策略中，一个分区对应的选择概率在整个的测试过程中是不变的，这一点可能不总是好的。因为引起故障的输入在输入域中趋向于聚簇在连续的区域[7-9]，也就是说存在一些分区更可能揭示软件中的故障。Cai 等人依据这一想法，利用软件的控制理论[10]提出了动态随机测试策略（DRT）[11]以改进传统的随机测

试与随机分区测试策略。软件的控制理论探索软件工程理论与控制理论相互作用的关系，被用来解决软件工程中的问题。DRT 策略的主要特点是在测试的过程中根据每一次测试用例的执行结果动态改变测试剖面：假设存在一个分区$c_i$，若该分区中的一个测试用例揭示了软件中的故障，那么认为该分区具有较高的故障检测能力，因此增大该分区被选择的概率，即$p_i + \varepsilon$。如果这个测试用例没有检测出故障，减小该分区被选择的概率，即$p_i - \delta$。

但是该策略仍然存在一些不足：

1. **找到高故障检测能力的分区的速度慢。**由于在测试过程中参数的取值很小，引起故障的输入所在的分区增加的概率幅度不明显，因此具有更高检测能力的分区很难在短时间内突显出来。特别是软件输入域的失效率很小的情况下，大多数的输入不能引起软件中的故障，因此能够造成故障的输入所在的分区难以保持较高的被选择概率。
2. **不同分区的测试结果相互影响。**某个分区被选择的概率受其它分区测试结果的影响,使得该分区被选择的概率无法准确地反映该分区真正的故障检测能力。将软件的输入域按照一定的方式划分为若干分区之后，每个分区的故障检测能力是独立的。但是在传统的 DRT 算法中每个分区的故障检测能力，随着其它分区的测试用例检测结果发生改变。
3. **每次调整概率的幅度相同。**直觉上，每次调整分区的概率幅度应当根据当前分区的概率大小，而不应该是一致的：如果当前分区被选择的概率比较大说明该分区在理论上具有更高的故障检测能力，但是并不能保证该分区的每一个测试用例都能揭示软件的故障，因此该分区的测试用例没有揭示故障时的调整幅度应当比被选择的概率较小分区没有检测出故障时的调整幅度要小。
4. **分区数目对测试策略有影响。**在黑盒测试中，同一个项目的不同分区策略以及同一个项目的相同分区策略，分区的数目也可能不同，不同数目的分区导致算法的检测效率发生改变。
5. **初始剖面对测试策略有影响。**一般情况下，初始测试剖面$\{p_1, p_2, \dots, p_m\}$中$p_1 = p_2 = \cdots = p_m$，即初始条件下每一个分区被选择的概率是均等的。当软件的输入域失效率高时，即存在很多测试用例能够造成软件故障，此时每一个子分区的故障检测能力可能相差不大。但是当软件的输入域失效率很低时，即只有很少的测试用例能够揭示故障中的故障，此时子分区之间的故障检测能力很有可能相差很大，甚至一些分区不具备故障检测能力。因此本文提出一种根据子分区测试用例数目占全部测试用例

的百分比作为初始剖面的概率分布，比较两种情况下，哪一种初始概率分布具有较高的故障检测率。

因此，加速找到具有高故障检测率的分区是一个很自然的想法弥补 DRT 策略在 1 方面存在的不足。本文通过为每一个分区绑定奖励因子与惩罚因子，并根据测试对象本身信息设定惩罚上限，如果某一个分区中的测试用例揭示了软件中的故障，那么该分区的奖励因子自增，惩罚因子设置为 0 并且下次依然在该分区中随机选择测试用例直到没有揭示软件中的故障，然后根据奖励因子确定该分区被选择概率的增长幅度。当某一个分区中的测试用例没有揭示软件中的故障时，该分区绑定的惩罚因子自增，然后调整测试剖面选择下一个分区。如果某一个分区的惩罚因子大于或者等于惩罚上限，意味着该分区具有很小的故障检测率，甚至不具备故障检测能力，因此将该分区被选择的概率记为 0。将这种测试策略命名为 RAPT。

为了弥补 2、3 方面的不足，本文用 Markov 链的状态转移矩阵思想，将分区作为状态，选择测试用例并执行当作在该状态下的行动，那么根据在某一个状态下测试用例执行情况调整转移到其它状态的概率。由此某一个分区被选择的概率只受该分区内测试用例执行结果的影响，不受其它分区测试用例执行结果的影响。并且在设计根据测试用例的执行结果调整分区被选择的概率时，本文对具体算法进行改进使得被选择概率大的分区没有揭示软件中的故障时概率调整幅度小，揭示故障时概率调整幅度大；被选择概率小的分区没有揭示软件中的故障时概率调整的幅度大，揭示故障时概率调整的幅度小。将这种测试策略命名为 MAPT 策略。

针对问题 4 本文根据五个程序的规格说明运用等价类划分法得到一种分区方式，然后将某一个分区再进行更细粒度的划分，得到分区数目更多的另一种分区方式。根据实验结果对比不同分区数目对各个测试策略的影响。

针对问题 5 本文采取均等的概率分布和不均等的概率分布作为初始剖面。不均等的初始概率分布的设置方式是根据每一个分区内的测试用例数目占输入域所有测试用例总数的百分比作为初始条件下某一分区被选择的概率。

本文通过实验发现在较多分区数目以及初始剖面为均等的概率分布时，MAPT、RAPT 策略的测试效率比 DRT、RPT 策略的效率高。

文章接下里的组织方式是：第二部分展示了相关工作。第三部分介绍了 MAPT 策略、RAPT 策略。第四部分展示了实验设置以及实验结果。数据分析和讨论在第五部分展示。第六部分展示了实验结论以及将来的工作。

## 2 相关工作

Weyuker[12]经过研究之后发现：分区测试可能是一个卓越的测试策略也可能是一个低效率的测试策略，分区测试的效率很大程度上取决于如何将产生错误输出的输入集中在某个或者某些分区中。Hamlet[3]认为成功的分区测试不能激发测试人员对软件质量的信心。Chen[4]认为当存在较高失效率的分区时，分区测试具有较高的检测能力。

Cai 结合了随机测试和分区测试的特点提出了分区随机测试（RPT）策略。RPT策略首先根据测试剖面选择分区$c_i$，然后在$c_i$中随机选择测试用例。Cai[10]利用软件控制理论提出了适应性测试策略(AT)，该策略的测试效率相对于随机测试、分区测试有很高的改进[5,10]。但是 AT 策略在实际中需要消耗大量的时间。为了解决这一问题 Cai 提出了动态随机测试策略(DRT)。在 DRT 策略中，测试剖面根据测试的反馈信息动态地改变。这里将完整的算法策略展示如下。

**步骤 1**：待测软件的输入域划分为m个不相交的分区：$C_1, C_2, \dots, C_m$，每一个分区中有$k_1, k_2, \dots, k_m$个测试用例。

**步骤 2**：初始化参数$\varepsilon, \delta$，并且$\varepsilon > 0, \delta > 0$。

**步骤 3**：根据每一个分区所对应的概率$p_i$随机选取一个分区$C_i$,在这里$p_1 + p_2 + \cdots + p_m = 1$。

**步骤 4**：随机地从$C_i$中挑选一个测试用例TC。

**步骤 5**：如果测试用例TC揭示了软件中的故障，就增大测试用例TC所在分区被选择的概率，同时减小其它分区被选择的概率，并把缺陷移除。

$$p_j = \begin{cases} p_j - \dfrac{\varepsilon}{m-1}, p_j \geq \dfrac{\varepsilon}{m-1} \\ 0 \qquad\qquad , p_j < \dfrac{\varepsilon}{m-1} \end{cases}$$

$$p_i = 1 - \sum_{j \neq i} p_j$$

**步骤 6**：如果测试用例TC没有找到缺陷，就减少$c_i$被选中的概率$p_i$，同时增大其它分区被选中的概率$p_j$。

$$p_i = \begin{cases} p_i - \delta, p_i \geq \delta \\ 0 \qquad , p_i < \delta \end{cases}$$

$$p_j = \begin{cases} p_j + \dfrac{\delta}{m-1}, p_i < \delta \\ p_j + \dfrac{p_i}{m-1}, p_i < \delta \end{cases}$$

**步骤 7：**检查停止条件，如果不满足，则跳转执行步骤 3，如果满足则停止测试。

从 DRT 策略的具体算法中可以看出，参数影响 DRT 策略的测试效率。Lv 在[13]中假设软件输入域的失效率已知、各个分区的失效率已知、测试过程中各个分区失效率保持不变以及测试用例执行之后放回原来的分区之中，通过理论分析的方式得到了$\varepsilon/\delta$的最佳取值范围。然而实际中很难知道输入域的失效率以及各个分区的失效率大小。Yang 在[14]中通过在实验的过程中统计每一个分区的成功检测率，然后调整$\varepsilon/\delta$的取值，该策略在软件不存在“难”检测的故障时，效果比较好。Li 提出了理论上的最佳测试剖面[17]，在测试过程中满足预先定义的标准之后将测试剖面转换成理论最佳的测试剖面。

# 3 基于 Markov 链的动态随机测试和基于奖惩机制的动态随机测试

这个章节主要介绍基于 Markov 链的动态随机测试(MAPT)策略以及基于奖惩机制的动态随机测试(RAPT)策略。

## 3.1 MAPT 策略

为了提高传统 DRT 策略的测试效率，MAPT 策略结合了传统随机算法与分区算法的特点，并引入软件的控制理论和 Markov 链的状态转移理论。

Markov 链具备“无后效应”，即要确定过程将来的状态，知道它此刻的情况就够了，并不需要对它以往状况的认识。对于有限个或可列个值$E_1, E_2, \cdots, E_n$，以$\{1,2,\cdots,\mathrm{n}\}$来标记$E_1, E_2, \cdots, E_n$并称它们为过程的状态，对于任意的$n \geq 0$及状态$i_1, i_2, \cdots, i_{n-1}, i, j$有：$\mathrm{P}\{X_{n+1} = \mathrm{j} | X_1 = i_1, X_2 = i_2, \cdots, X_n = i_n\} = \mathrm{P}\{X_{n+1} = \mathrm{j} | X_n = \mathrm{i}\}$

因此一旦 Markov 链的初始分布$\mathrm{P}\{X_0 = i_0\}$给定，其统计特性完全由条件概率$\mathrm{P}\{X_n = i_n | X_{n-1} = i_{n-1}\}$决定。为了接下来的讨论假设状态空间$\mathrm{S} = \{1,2,\ldots,\mathrm{m}\}$。

定义 1（转移概率）：条件概率$\mathrm{P}\{X_{n+1} = \mathrm{j} | X_n = \mathrm{i}\}$为 Markov 链的一步转移概率，简称转移概率。

定义 2（时齐 Markov 链）：当 Markov 链的转移概率$\mathrm{P}\{X_{n+1} = \mathrm{j} | X_n = \mathrm{i}\}$只与状态 i,j有关，而与n无关时，称 Markov 链为时齐的，并记为$p_{ij} = P\{X_{n+1} = j | X_n = i\}(n \geq 0)$。

由定义 2 可以将$p_{ij}(i, j \in S)$排成一个矩阵的形式，令

$$P = (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

则称P为转移矩阵。

假设软件的输入域有k个测试用例，被划分到m个不相交的分区$C_1, C_2, \ldots, C_m$中并且每一个分区有$k_i$个测试用例，$k_1 + k_2 + \cdots + k_m = k$。如果将每个时间点$t(t = 0,1,2,\cdots)$测试用例所在的分区作为时刻 t 测试系统所处的状态，则整个状态空间为$S = \{s_t, t \geq 0\} = \{C_1, C_2, \ldots, C_m\}$。根据在$s_t = C_i$状态下测试用例的执行结果可以计算调整到状态$s_{t+1} = C_j$的转移概率。如果$s_t = C_i$状态下检测出软件中存在缺陷，那么就增大下一时刻转移到$s_{t+1} = C_i$的概率，同时减小转移到其它状态的概率；反之，如果在$s_t = C_i$状态下没有检测出软件中存在缺陷那么就减少下一步转移到$s_{t+1} = C_i$状态的概率，同时增大转移到其它状态的概率。另外，根据当前状态下测试用例的执行结果每次增大或者减少的转移概率的幅度应该是不同的。一般情况下，一方面即便某一个分区具有较强的检测出软件存在故障的能力，也不可能每一次在该分区中选择测试用例都会检测出故障。另一方面如果某一个分区被选择的概率比较大说明在以往的测试过程中较多的检测出了软件中的缺陷或者理论上有较大的可能揭示软件中的故障。因此在测试过程中增大或者减少某一分区被选择的概率幅度时应当与该分区当前被选择概率有关：如果当前分区被选择的概率较大，那么增大该分区被选择的概率幅度就越大，减小该分区被选择的幅度就越小。整个过程的状态转移可以用转移矩阵表示。在整个软件测试过程中，将每一个时间点选取测试用例并执行作为一次决策行动，则行动全体组成整个行动空间$A = \{a_t, t \geq 0\} = \{1,2,\cdots,k\}$，并且每个时间点的状态$S_t$和所采取的行动都会影响到下一个时间点$t + 1$的状态$S_{t+1}$。因此，整个测试过程形成一个 Markov 决策过程。

MAPT 算法描述了 MAPT 测试策略的框架。为了方便表示，不妨设状态空间$S = \{1,2,\ldots,m\}$。其输入包括程序转移概率矩阵P，算法的参数γ,τ，停止条件

MART 算法的具体过程：

开始时刻初始化测试剖面$p'$，则初始转移矩阵$P = p_{ij} = (p', p', \ldots, p')'$其中 $(i = 1,2,\ldots,m; j = 1,2,\ldots,m)$。

步骤 1：根据当前分区到其它分区所对应的转移概率$p_{ij}$随机选取一个分区$C_i$(第一次根据测试剖面选择分区)，在这里$p_{i1} + p_{i2} + \cdots + p_{im} = 1$。转步骤 2。

步骤 2：等概率随机地从分区$C_i$中选取一个测试用例 TC,转步骤 3。

步骤 3：执行选中的测试用例 TC：

如果测试用例 TC 揭示了软件中的故障，就增大测试用例 TC 所在状态转移到本身的概率$p_{ii}$，同时减小转移到其它状态的概率$p_{ij}(i \neq j)$，并把缺陷移除：

$$p_{\mathrm{ij}} = \begin{cases} p_{ij} - \dfrac{\gamma \times p_{ii}}{m-1}, & p_{ij} \geq \dfrac{\gamma \times p_{ii}}{m-1} \\ p_{ij}, & p_{ij} < \dfrac{\gamma \times p_{ii}}{m-1} \end{cases} \quad (i \neq j)$$

$$p_{ii} = 1 - \sum_{i \neq j} p_{ij}$$

如果测试用例 TC 没有找到缺陷，就减少测试用例 TC 所在状态转移到本身的概率$p_{ii}$，同时增大转移到其它状态的概率$p_{ij}(i \neq j)$：

$$p_{ij} = \begin{cases} p_{ij} + \dfrac{\tau \times p_{ij}}{m-1}, & p_{ii} > \dfrac{\tau \times (1-p_{ii})}{m-1} \\ p_{ij}, & p_{ii} < \dfrac{\tau \times (1-p_{ii})}{m-1} \end{cases}$$

$$p_{ii} = \begin{cases} p_{ii} - \dfrac{\tau \times (1-p_{ii})}{m-1}, & p_{ii} > \dfrac{\tau \times (1-p_{ii})}{m-1} \\ p_{ii}, & p_{ii} > \dfrac{\tau \times (1-p_{ii})}{m-1} \end{cases}$$

更新转移矩阵：

$$\begin{pmatrix} * & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ p_{i1} & \cdots & p_{ii} & \cdots & p_{im} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ * & * & * & \cdots & * \end{pmatrix}$$

步骤 4：检查测试停止条件，如果不满足则转步骤 1；如果满足则停止测试。

MAPT 算法

Input: γ,τ,Condition,$p$;

Output: Report

1: Initialize P

2: While !Condition do

3: If i = 0 then

4: $P_{ij} = p$;

5: i = ChooseState($P_{ij}$);

6: Else

7: i = ChooseState($P_{ij}$)

8: End If

9: TC = ChooseTestCase(i);

10: If diff (result(TC),expectantresult(TC)) then

11: RecordReport(TC);

12: increase $P_{ii}$ && decrease $P_{ij}$

13: Else

14: RecordReport(TC);

15: decrease $P_{ii}$ && increase $P_{ij}$

16: End If

17: End While

Condition,初始的测试剖面 p,某一时刻测试系统所处的状态i(测试之前$i = 0$)以及状态j($i, j \in S$)。其输出是每一个测试用例的执行结果形成的报告。一般情况下，系统检测到一个故障，测试了一段既定的时间以及揭示一定数目的故障均可作为停止条件。本文将揭示软件中第一个故障作为测试策略的停止条件。算法分为两个步骤。

步骤一，挑选测试用例并执行：测试开始时先根据测试剖面确定下一时刻所处的状态i，并且状态转移矩阵根据初始测试剖面进行初始化$P = \{p, p, \ldots, p\}'$。然后在状态i对应的分区之中随机选择测试用例 TC 并在程序中执行，然后比较结果与该测试用例预期的结果是否一致。

步骤二，将测试用例 TC 的测试结果记录在 Report 中，并且如果该测试用例的执行结果与预期输出不一致，即揭示了软件中的故障，则增大下一时刻仍处于状态i的概率，即增大$P_{ii}$的值，减小$P_{ij}$的值：在根据测试结果调整状态i到状态j($j \neq i$)的转移概率$P_{ij}$时，如果当前的概率$P_{ij} < \gamma * P_{ii}/(m-1)$则$P_{ij}$已经足够小，因此不改变$P_{ij}$的值，否则$P_{ij} = P_{ij} - \gamma * P_{ii}/(m-1)$,由此得到下一个状态为i的概率为$P_{ii} = 1 - \sum_{j \neq i} p_{ij}$,所以当目前状态i被选择的概率大时，下一步调整转移到其它状态的概率减小的幅度就相对较大，从而使故障检测能力较高的分区具有更大的机会被选择到。如果该测试用例的执行结果与预期结果一致，则减小下一时刻仍

在状态i的概率$P_{ii}$,增大转移到其它状态的概率$P_{ij}$。如果$P_{ii} < \tau * (1 - P_{ii})/(m - 1)$不等式成立,则说明下一步转移到状态i的概率已经非常小此时$P_{ii} = P_{ii}, P_{ij} = P_{ij}$。如果不等式不成立则增大转移到状态j(j ≠ i)的概率$P_{ij} = P_{ij} + \tau * P_{ij}$/(m-1), 减小转移到状态i的概率$P_{ii} = P_{ii} - \tau * (1 - P_{ii})/(m - 1)$。如果当前状态i被选择的概率比较大则$\tau$ * (1 - $P_{ii}$)/(m-1)的值就相对小，因此测试用例没有揭示故障时，当目前状态 i 被选择的概率较大时，下一步仍然转移到状态 i 的概率减少的幅度越小。增大转移到状态j(j ≠ i)的概率时，增加的幅度跟当前状态j被选择概率的大小有关，概率越大，增加的幅度越大，概率越小增加的幅度越小。

## 3.2 RAPT 策略

在软件中不存在难检测的故障时，Yang 在[16]中提出的 A-DRT 策略的测试效率比传统的 DRT 策略有明显的提高；但是软件存在难检测的故障时，效果不理想。当软件的失效率高时，软件内的缺陷很多测试策略都能用较小的代价揭示出来。但是当软件的失效率低时，不同检测策略的效率差别很大。在以往的测试活动中发现，当失效率很低时 DRT 策略的测试效率相对于 RT 策略没有提高或者提高不明显。直觉地，引起故障的输入在输入域中趋向于聚簇在连续的区域，即存在一个或者少数分区具有较高的检测能力。因此在软件输入域的失效率较低时，往往一些分区内不具备揭示软件中缺陷的能力或者具备较小的检测能力。另一方面由于每次调整概率的幅度很小，并且某一个分区被选择的概率易受到其它分区的测试结果的影响，使得那些不具备或者具备很小的检测能力的分区仍然被不断的选择，最终具有较高检测能力的分区不容易在短时间内突显出来.因此 DRT 策略在软件输入域的失效率低时，测试效率不高。为了缓解这一问题，本文提出了基于奖惩机制的动态随机测试策略(RAPT)，该策略旨在加速测试的过程：如果分区$C_i$内的测试用例揭示了软件中的缺陷，下一次仍在该分区内选择测试用例并且该分区绑定的奖励因子自增一次，对应的惩罚因子清 0，直到该分区中的测试用例没有检测出软件中的缺陷，奖励因子清 0，惩罚因子+1。奖励因子越大该分区对应的概率增加的越多。相反地，如果存在这样一个分区：累计 n 次选中该分区，但是该分区中的测试用例均没有揭示出软件中存在缺陷，那么就认为该分区具有较低的检测能力，甚至不具备检测能力，让该分区对应的选择概率为 0。

RART 算法的具体过程：

假定软件测试的输入域中的测试用例划分到 m 个不相交的分区中，输入域中共

有k个测试用例，用$C_1, C_2, \ldots, C_m$来表示这 m 个分区，每一个分区有 $k_i$个测试用例。初始化每个分区的奖励因子$reward_i = 0$，惩罚因子$punishment_i = 0$，惩罚上限$boundary_i$。

步骤 1：根据当前各个分区所对应的概率$p_i$选取分区$C_i$,其中$p_1 + p_2 + \cdots + p_m = 1$。

步骤 2：等概率随机地从分区$C_i$中选取一个测试用例 TC。

步骤 3：执行选中的测试用例 TC。如果测试用例 TC 揭示了软件中的故障转步骤 4，反之转步骤 5。

步骤 4：分区$C_i$的奖励因子$reward_i = reward_i + 1$，惩罚因子$punishment_i = 0$，并移除缺陷。转步骤 2。

步骤 5：分区$C_i$的惩罚因子$punishment_i = punishment_i + 1$。如果$reward_i \neq 0$就增大测试用例 TC 所在分区对应的概率$p_i$,同时减小其它分区被选择的概率$p_j (i \neq j)$：

$$p_j = \begin{cases} p_j - \dfrac{(1 + \ln reward_i) \times \varepsilon}{m - 1}, & p_j \geq \dfrac{(1 + \ln reward_i) \times \varepsilon}{m - 1} \\ 0, & p_j < \dfrac{(1 + \ln reward_i) \times \varepsilon}{m - 1} \end{cases}$$

$$p_i = 1 - \sum_{i \neq j} p_j$$

如果$reward_i = 0$，就减少该分区对应的概率$p_i$，同时增大其它分区对应的概率$p_j (i \neq j)$，如果该分区的惩罚因子$punishment_i \geq boundary_i$则该分区对应的概率$p_i = 0$：

$$p_i = \begin{cases} p_i - \delta, & p_i \geq \delta \\ 0, & p_i < \delta \ \text{ or } \ punishment_i = boundary_i \end{cases}$$

$$p_j = \begin{cases} p_j + \dfrac{\delta}{m - 1}, & p_i \geq \delta \\ p_j + \dfrac{p_i}{m - 1}, & p_i < \delta \ \text{ or } \ punishment_i = boundary_i \end{cases}$$

步骤 6：检查测试停止条件，如果不满足则转步骤 1；如果满足则停止测试。

RAPT 算法

Input ε, δ, m, Condition, p, boundary, punishment = 0, $reward = 0$;

Output: Report

1: While !Condition do
2: i = ChoosePartition(p);
3: TestCase: TC = ChooseTestcase(i);
4: If diff (result(TC),expectantresult(TC)) then
5: RecordReport(TC);
6: $reward_i++$ && $punishment_i = 0$;
7: goto: TestCase;
8: Else
9: RecordReport(TC) && $\text{punishment}_i++$;
10: If $reward_i \neq 0$ then
11: increase $p_i$ && decrease $p_j$;
12: $reward_i = 0$;
13: Else
14: If $p_i \geq \delta$ && $\text{punishment}_i < boundary_i$ then
15: decrease $p_i$ && increase $p_j$;
16: Else
17: $p_i = 0$ && increase $p_j$;
18: End If
19: End If
20: End If
21: End While

RAPT 算法描述了 RAPT 测试策略的框架。为了表示方面假设待测软件的输入域划分为$\{C_1, C_2, \dots, C_m\}$m 个分区，每一个分区用$\{1,2,\dots,\text{m}\}$表示$(\text{i},\text{j} \in \{1,2,\dots,\text{m}\})$。其输入有参数 ε，δ，待测软件的分区数目m，停止条件Condition，初始测试剖面p，每一个分区挑选测试用例的上限$\text{boundary}_i$以及惩罚因子$\text{punishment}_i$和奖励因子$reward_i$。本文 RAPT 策略的停止条件Condition和 MAPT 策略的停止条件相同。每一个分区的对应的$\text{boundary}_i$应该根据具体的待测软件由有经验的测试人员设定。在测试之前，每一个分区对应的奖励因子以及惩罚因子都为 0。输出为一个测试报告，该报告包含了从测试开始到测试结束执行的测试用例的执行信息。该算法包含两个步骤。

第一步，根据测试剖面选择分区$i$然后在该分区中随机选择测试用例 TC 并在待测软件中执行。

第二步，对比测试用例 TC 的执行结果与预期输入是否一致。如果不一致，即 TC 揭示了软件中的故障，那么$i$分区的奖励因子$reward_i$自增，惩罚因子$\text{punishment}_i$清零。并且下一个测试用例仍然在$i$分区中选择，直到在$i$分区中选到

的测试用例没有揭示软件中的故障。接着$\text{punishment}_i$自增并且判断$reward_i$的值，如果$reward_i = 0$，$\text{punishment}_i < \text{boundary}_i$,则说明$i$分区中的测试用例没有揭示软件中的故障并且该分区连续没有揭示故障的测试用例数目没有到达上限，因此减少$\text{p}_i$增大$\text{p}_j(j \neq i)$(减小增大的机制同DRT策略)，并且$\text{punishment}_i$自增；若$reward_i = 0$并且$\text{punishment}_i = \text{boundary}_i$，则说明在n次的实验中，一共选到了$\text{boundary}_i$次$i$分区，但是该分区中的测试用例均没有揭示软件中的错误。由此该分区可能具有很低的故障检测能力甚至不具备故障检测能力，所以令$\text{p}_i = 0$并且增大$\text{p}_j(j \neq i)$( 增大的机制同 DRT 策略),最后令$\text{punishment}_i = 0$； 当$reward_i \neq 0$，即$i$分区连续有$reward_i$个测试用例揭示软件中的故障说明$i$分区可能具有较高的故障检测率，因此增大$\text{p}_i$减少$\text{p}_j(j \neq i)$。如果$\text{p}_j < \frac{(1+lnreward_i)\times\varepsilon}{m-1}$,则$\text{p}_j = 0$，否则$p_j - \frac{(1+\ln reward_i)\times\varepsilon}{m-1}$。并且$p_i = 1 - \sum_{i\neq j} p_j$。并且$reward_i$的值越大$\text{p}_i$增加的幅度就越大。最后令$reward_i = 0$，并根据更新后的测试剖面重新选择分区。

## 3.3 MAPT 算法以及 RAPT 算法复杂度分析

对于 MAPT 算法我们不妨设在一次软件测试过程中将软件的输入域分为 m 个分区。在理想情况下，执行一个测试用例就满足停止条件，此时执行的算法的语句数量为$\text{T(n)} = \text{m} + \text{k}$。其中 m 是更改转移矩阵某一行的数值需要执行的语句次数，k 是从根据测试剖面选择分区到选择测试用例并执行这一过程执行的可数的简单语句数目。由于在实际的情况中分区的个数总是有限的，因此这时算法的渐进时间复杂的为$\text{O}(1)$。如果不能一个测试用例就找出所有的缺陷，那么不妨设在整个测试的过程中共执行了 n 个测试用例，这时执行的语句数目为$\text{T(n)} = \text{n} * \text{m} + \text{n}$。当 n 很大时，这时的时间复杂度为$\text{O}(\text{n})$。

相似地，对 RAPT 算法，在理想情况下，执行一个测试用就满足了停止条件，这时的时间复杂度为$\text{O}(1)$。如果执行的测定用例的数目比较多时，时间复杂度为$\text{O}(\text{n})$。

## 3.4 Remarks

[10]中的实验可以看出增大$\varepsilon/\delta$的值可以提高 DRT 策略的效率，因此让$\varepsilon/\delta = r_M + (r_\Delta - r_M) * K(r_M = 1/\theta_M - 1, r_\Delta = 1/\theta_\Delta - 1$，并且 K 值为 0.8 时 DRT 策略具有更高的故障检测效率。保守地， ε 被设置一个相当小的数 ε $= 0.05$，本实验中$\theta_M$、$\theta_\Delta$是根据真实的故障检测率确定的大小。每一个实验对象，DRT 策略和 RAPT 策略的参数大小相同。所有的实验对象 MAPT 策略的参数均设置为 γ $=$ τ $= 0.1$。

RAPT 策略的每一个分区的惩罚上限根据具体实验对象的不同，各个分区的测试用例数目以及测试用例的特点设置的具体值也可能不同。本文由于植入的故障都相对难杀死，因此惩罚上限为每一个分区测试用例数目的 70%。

初始的测试剖面$\{p_1, p_2, \ldots, p_m\}$应该有测试工程师根据以往的测试经验设定。本文采取两种方式设定初始测试剖面。第一种方式为$p_1 = p_2 = \cdots = p_m = 1/m$；第二种设置方式为$p_i = k_i/k$，$k_i$表示$C_i$分区内的测试用例数目，k表示 SUT 输入域的测试用例总数，下文提到的不均等的概率分布作为初始剖面均是指这种概率分布。这两种初始化测试剖面的方式应用于 RPT、DRT、MAPT、RAPT 四种测试策略中。

# 4 实验设置

为了验证 MAPT、RAPT 测试策略比 DRT、RPT 测试策略据具有更高的故障检测效率，本文用以上四种测试策略对三个真实的程序进行测试，用两种度量指标量化每一种测试策略的效率。并且划分 3 种不同数目的分区，设置均等和不均等的初始概率分布作为测试剖面，探究分区数目以及初始测试剖面对 RPT、DRT、MAPT、RAPT 测试策略的测试效率的影响。

## 4.1 实验对象

为了避免测试策略在特定的程序中具有更高的故障检测能力，本文在 Software artifact Infrastructure Repository(SIR)网站下载了三个真实的程序，每一个程序附带测试用例集以及故障。为了模拟实际的软件测试，本文选取的程序代码行数均大于 5K。在 SIR 网站中，每一个程序都有不同的版本以及对应的测试用例集和故障。实验对象的基础信息展示在表 4.1，其中 gzip 程序的 V3 版本所选择的测试用例集检测不出故障，因此本文没有对该程序测试。

表 4.1 实验对象信息

| Subject | LOC | Versions | Test Cases | Faults | Partitions |
|---|---|---|---|---|---|
| grep | 10068 | V1 | 470 | 2 | {3,9,13} |
| | | V2 | | 2 | {3,9,13} |
| | | V3 | | 2 | {3,9,13} |
| | | V4 | | 1 | {3,9,13} |
| gzip | 5680 | V1 | 214 | 3 | {4,6,12} |
| | | V2 | | 1 | {4,6,12} |
| | | V4 | | 2 | {4,6,12} |
| | | V5 | | 3 | {4,6,12} |
| make | 35545 | V1 | 793 | 2 | {4,8,16} |
| | | V2 | | 1 | {4,8,16} |

## 4.2 实验设计

TSL[15]是用来书写测试规格说明书的语言。基于测试规格说明书中的信息可以产生大量的测试帧。为每一个测试帧中的 choice 指定一个具体的值，可以得到一个具体的测试用例。本文的分区方式是考虑测试规格说明书中的不同数目的 categories，得到不同粒度的分区方式，每一个实验对象的分区数目展示在表 4.1 的最后一列：

1. 粗粒度：选取测试规格说明书中的一个 category，这个 category 的 choice 应当具有较少的约束，或者没有约束以便与其它 categories 中的 choice 进行组合。然后将该 category 中的每一个 choice 作为一个分区，得到粗粒度的分区方式。对于一些 categories 以及这些 categories 所涉及的功能，测试人员在编辑测试脚本时可能认为这些功能比较简单不容易出错或者根据以往的测试经验，用很少的测试用例对这些功能进行测试就能增加人们对这些功能的信心。因此这这些 categories 中的 choice 加上 single 或者 error 约束条件，不与其它 categories 中的 choices 相互联系，并且只产生一个测试用例。因此本文在划分分区时不考虑这样的 categories。
2. 中等粒度：考虑 2 个或者多个 categories,并且这些 categories 中的 choices 有较少的约束，或者没有约束。然后将不同 categories 中的 choice 根据测试规格说明书进行组合得到分区数目比粗粒度分区数目更多的防区方式。
3. 细粒度：考虑几乎所有的 choices 没有约束或者约束较少的 categories，然后将这些 categories 中的 choices 进行组合，得到数目更多的分区方式。

对于每一个实验对象，本实验先用随机测试策略对每一个故障进行测试，重复 20

遍得到每一个故障被检测到时用的测试用例数目的平均值。比较每一个故障用到的测试用例数目然后取相对难杀死的故障(检测到该故障用到的测试用例数目较多)，每一个实验对象测试的故障数目展示在表 4.1 的倒数第二列。

本实验检测了四个策略：随机分区测试策略(RPT)，动态随机测试策略(DRT)，以及本文提到的基于 Markov 链的动态随机测试策略(MAPT)和基于奖惩机制的动态随机测试策略(RAPT)。在测试过程中，如果一个故障被检测到就立即移除该故障。测试的停止条件为所有的故障都被揭示和移除。

由于计算机产生的随机数是伪随机,如果不指定随机数种子，计算机将以当前时间为种子随机产生随机数。如果这样做一方面导致实验不可重复，另一方面也使得不同测试策略的差异是随机数产生的还是策略本身产生的无法确定。因此对于同一个实验对象 20 次重复实验的随机数种子为$\{1,2,\ldots,20\}$。

为了反映四个测试策略的故障检测效率，在实验中运用了 3 个度量标准：

1. F，揭示第一个故障需要的测试用例数目。F的均值用$\bar{F}$表示。
2. NF，揭示第一个故障之后揭示第二个故障需要的测试用例数目。NF的均值用$\overline{NF}$表示。
3. SD，反应了不同测试策略在 F 度量标准下的稳定性。F 度量标准的标准差用SD_F表示，NF 度量标准的标准差用SD_NF表示。

接下来的章节展示了不同实验的各个度量标准的值。本文没有考虑执行每一个测试用例花费的代价，假设同一个实验中，每一个测试用例花费的代价是相同的。在这种情况下，F度量标准和NF度量标准可以用来比较不同测试策略检测故障的效率，SD_F和SD_NF可以反应测试策略在F度量标准和NF度量标准下的稳定性。

表 4.2 和表 4.3 分别展示了在不同实验对象下，RPT、DRT、MAPT、RAPT 策略F度量标准以及NF度量标准的测试结果的平均值和标准差以及 DRT、MAPT、RAPT 策略的相对于 RPT 策略效率的提升率。在表 4.2 和表 4.3 中 initial profile 指的是测试对象的分区设置以及初始剖面设置用符号$\tilde{x}$或者$x$表示。其中$\tilde{x}$表示分区的数目为$\tilde{x}$，采用不均等的初始概率分布作为测试剖面；$\tilde{x}$表示分区的数目为$x$，采用均等的初始概率分布作为测试剖面。例如 grepV1 实验中$\tilde{3}$表示实验对象为 grep 程序的第一个版本，分区数目为 3，采用的是均等的初始概率作为初始测试剖

## 4.2(a) grep 程序 F 度量标准实验结果

| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | SD_F | F | SD_F | improvement rate | F | SD_F | improvement rate | F | SD_F | improvement rate |
| grepV1 | $\tilde{3}$ | 231.70 | 199.21 | 258.90 | 254.89 | -11.74% | 217.9 | 170.92 | 5.96% | 159.10 | 127.63 | 31.33% |
| | $\bar{3}$ | 123.70 | 117.16 | 156.90 | 146.30 | -26.84% | 118.65 | 85.66 | 4.08% | 121.50 | 130.49 | 1.78% |
| | $\tilde{9}$ | 214.90 | 153.41 | 179.15 | 169.53 | 16.64% | 164.7 | 174.15 | 23.36% | 129.70 | 136.35 | 39.65% |
| | $\bar{9}$ | 169.45 | 191.53 | 160.20 | 125.45 | 5.46% | 153.85 | 139.26 | 9.21% | 148.10 | 116.02 | 12.60% |
| | $\widehat{13}$ | 188.65 | 193.18 | 181.45 | 202.65 | 3.82% | 175.85 | 160.6 | 6.79% | 117.45 | 96.21 | 37.74% |
| | $\overline{13}$ | 150.15 | 115.33 | 160.8 | 320.02 | -7.09% | 124.9 | 166 | 16.82% | 144.35 | 167.3 | 3.86% |
| grepV2 | $\tilde{3}$ | 18.15 | 14.89 | 18.00 | 12.53 | 0.83% | 17.15 | 11.79 | 5.51% | 14.30 | 10.33 | 21.21% |
| | $\bar{3}$ | 13.70 | 9.97 | 13.60 | 12.61 | 0.73% | 11.6 | 9.11 | 15.33% | 10.60 | 8.44 | 22.63% |
| | $\tilde{9}$ | 7.65 | 5.15 | 6.55 | 7.69 | 14.38% | 5.5 | 4.66 | 28.10% | 4.90 | 4.01 | 35.95% |
| | $\bar{9}$ | 12.85 | 13.46 | 17.15 | 13.83 | -33.46% | 12.3 | 10.45 | 4.28% | 9.30 | 9.6 | 27.63% |
| | $\widehat{13}$ | 8.70 | 9.57 | 7.40 | 6.03 | 14.94% | 6.7 | 6.38 | 22.99% | 5.80 | 4.16 | 33.33% |
| | $\overline{13}$ | 16.70 | 13.19 | 15.95 | 13.04 | 4.49% | 10.4 | 9.52 | 37.72% | 12.35 | 11.36 | 26.05% |
| grepV3 | $\tilde{3}$ | 51.20 | 53.06 | 41.40 | 39.60 | 19.14% | 26 | 23.45 | 49.22% | 22.40 | 22.68 | 56.25% |
| | $\bar{3}$ | 35.00 | 33.02 | 28.30 | 20.11 | 19.14% | 25.95 | 21.73 | 25.86% | 22.45 | 20.66 | 35.86% |
| | $\tilde{9}$ | 65.80 | 44.71 | 90.85 | 80.78 | -38.07% | 87.6 | 107.43 | -33.13% | 36.05 | 29.97 | 45.21% |
| | $\bar{9}$ | 36.45 | 25.22 | 36.00 | 31.98 | 1.23% | 28.55 | 21.77 | 21.67% | 25.45 | 26 | 30.18% |
| | $\widehat{13}$ | 72.95 | 97.52 | 72.00 | 67.60 | 1.30% | 54.85 | 39.84 | 24.81% | 43.75 | 31.16 | 40.03% |
| | $\overline{13}$ | 26.90 | 25.76 | 25.60 | 30.75 | 4.83% | 24.85 | 22.08 | 7.62% | 23.50 | 28.27 | 12.64% |
| grepV4 | $\tilde{3}$ | 309.50 | 319.26 | 463.20 | 357.67 | -49.66% | 267 | 217.17 | 13.73% | 223.20 | 183.55 | 27.88% |
| | $\bar{3}$ | 246.00 | 158.27 | 264.40 | 181.92 | -7.48% | 235.2 | 197.87 | 4.39% | 209.65 | 159.55 | 14.78% |
| | $\tilde{9}$ | 189.45 | 182.10 | 187.75 | 196.36 | 0.90% | 148.6 | 153.72 | 21.56% | 213.00 | 164.04 | -12.43% |
| | $\bar{9}$ | 283.20 | 239.73 | 200.55 | 162.13 | 29.18% | 143.85 | 137.57 | 49.21% | 193.80 | 165.31 | 31.57% |
| | $\widehat{13}$ | 224.00 | 202.29 | 222.40 | 194.50 | 0.71% | 209.25 | 192.89 | 6.58% | 183.90 | 158.04 | 17.90% |
| | $\overline{13}$ | 279.05 | 357.56 | 305.85 | 245.69 | -9.60% | 207.75 | 177.93 | 25.55% | 226.35 | 328.55 | 18.89% |

## 4.2(b) make 程序 F 度量标准实验结果

| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | SD_F | F | SD_F | improvement rate | F | SD_F | improvement rate | F | SD_F | improvement rate |
| makeV1 | $\tilde{4}$ | 98.25 | 88.44 | 89.2 | 65.48 | 9.21% | 79.70 | 75.26 | 18.88% | 79.95 | 68.34 | 18.63% |
| | $\bar{4}$ | 89.25 | 73.95 | 86.4 | 89.09 | 3.19% | 86.10 | 75.82 | 3.53% | 73.05 | 65.15 | 18.15% |
| | $\tilde{6}$ | 127.40 | 127.74 | 122.55 | 105.90 | 3.81% | 117.25 | 86.86 | 7.97% | 104.10 | 111.69 | 18.29% |
| | $\bar{6}$ | 134.60 | 92.97 | 122.55 | 105.90 | 8.95% | 101.85 | 107.9 | 24.33% | 117.75 | 121.42 | 12.52% |
| | $\widehat{12}$ | 222.20 | 240.23 | 191.85 | 177.48 | 13.66% | 183.80 | 177.7 | 17.28% | 177.65 | 166.27 | 20.05% |
| | $\overline{12}$ | 134.30 | 140.02 | 129.9 | 130.02 | 3.28% | 104.75 | 90.18 | 22.00% | 102.4 | 97.19 | 23.75% |
| makeV2 | $\tilde{4}$ | 458.70 | 513.41 | 402.65 | 415.83 | 12.22% | 370.35 | 352.9 | 19.26% | 362.65 | 436.23 | 20.94% |
| | $\bar{4}$ | 413.15 | 541.83 | 388.7 | 558.38 | 5.92% | 350.25 | 400.9 | 15.22% | 385.65 | 473.16 | 6.66% |
| | $\tilde{6}$ | 557.60 | 376.94 | 431.2 | 386.37 | 22.67% | 413.65 | 331.4 | 25.82% | 405.85 | 241.42 | 27.21% |
| | $\bar{6}$ | 272.90 | 289.02 | 330.1 | 271.90 | -20.96% | 267.50 | 223.4 | 1.98% | 225.65 | 190.86 | 17.31% |
| | $\widehat{12}$ | 652.25 | 575.37 | 644.2 | 628.98 | 1.23% | 622.55 | 741.1 | 4.55% | 603.80 | 585.59 | 7.43% |
| | $\overline{12}$ | 456.55 | 488.25 | 324.25 | 325.20 | 28.98% | 317.85 | 311.1 | 30.38% | 318.55 | 329.4 | 30.23% |

## 4.2(c) gzip 程序 F 度量标准实验结果

| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | SD_F | F | SD_F | improvement rate | F | SD_F | improvement rate | F | SD_F | improvement rate |
| gzipV1 | $\tilde{4}$ | 38.55 | 28.89 | 34.05 | 38.34 | 11.67% | 24.25 | 18.2 | 37.09% | 21.90 | 19.17 | 43.19% |
| | $\bar{4}$ | 98.60 | 73.95 | 49.80 | 50.54 | 49.49% | 39.70 | 24.79 | 59.74% | 76.40 | 64.7 | 22.52% |
| | $\tilde{6}$ | 60.85 | 42.83 | 64.05 | 76.87 | -5.26% | 43.20 | 33.36 | 29.01% | 46.10 | 59.34 | 24.24% |
| | $\bar{6}$ | 86.35 | 89.39 | 66.95 | 68.23 | 22.47% | 66.15 | 46.99 | 23.39% | 65.25 | 62.81 | 24.44% |
| | $\widetilde{12}$ | 107.50 | 111.65 | 91.20 | 63.18 | 15.16% | 76.30 | 69.61 | 29.02% | 61.65 | 50.4 | 42.65% |
| | $\overline{12}$ | 117.75 | 106.82 | 119.05 | 145.92 | -1.10% | 111.70 | 117.9 | 5.14% | 70.45 | 67.68 | 40.17% |
| gzipV2 | $\tilde{4}$ | 66.50 | 70.50 | 56.70 | 47.66 | 14.74% | 54.85 | 38.7 | 17.52% | 36.20 | 33.68 | 45.56% |
| | $\bar{4}$ | 21.80 | 13.11 | 16.90 | 14.39 | 22.48% | 19.10 | 17.68 | 12.39% | 11.25 | 9.04 | 48.39% |
| | $\tilde{6}$ | 51.25 | 38.94 | 43.45 | 56.93 | 15.22% | 42.95 | 34.47 | 16.20% | 44.65 | 34.71 | 12.88% |
| | $\bar{6}$ | 19.10 | 18.23 | 17.55 | 19.96 | 8.12% | 12.25 | 12.36 | 35.86% | 15.75 | 15.29 | 17.54% |
| | $\widetilde{12}$ | 40.55 | 27.01 | 28.70 | 25.81 | 29.22% | 28.10 | 24.64 | 30.70% | 27.80 | 29.08 | 31.44% |
| | $\overline{12}$ | 17.15 | 13.25 | 12.15 | 12.95 | 29.15% | 10.90 | 8.5 | 36.44% | 13.15 | 13.63 | 23.32% |
| gzipV4 | $\tilde{4}$ | 40.10 | 29.76 | 34.35 | 24.90 | 14.34% | 34.35 | 28.11 | 14.34% | 21.40 | 16.42 | 46.63% |
| | $\bar{4}$ | 70.65 | 62.68 | 58.45 | 42.09 | 17.27% | 47.70 | 30.49 | 32.48% | 58.40 | 45.61 | 17.34% |
| | $\tilde{6}$ | 63.55 | 57.17 | 51.80 | 58.66 | 18.49% | 51.45 | 34.6 | 19.04% | 45.30 | 44.18 | 28.72% |
| | $\bar{6}$ | 107.55 | 69.83 | 76.05 | 61.10 | 29.29% | 63.65 | 81.36 | 40.82% | 74.75 | 55.76 | 30.50% |
| | $\widetilde{12}$ | 136.60 | 97.89 | 131.05 | 121.41 | 4.06% | 101.05 | 65.72 | 26.02% | 111.60 | 103.42 | 18.30% |
| | $\overline{12}$ | 99.80 | 101.66 | 91.25 | 100.34 | 8.57% | 68.60 | 82.48 | 31.26% | 81.05 | 73.99 | 18.79% |
| gzipV5 | $\tilde{4}$ | 5.95 | 5.86 | 5.00 | 3.34 | 15.97% | 4.75 | 3.49 | 20.17% | 4.05 | 2.99 | 31.93% |
| | $\bar{4}$ | 46.15 | 44.76 | 13.00 | 7.31 | 71.83% | 22.40 | 17.18 | 51.46% | 12.05 | 8.06 | 73.89% |
| | $\tilde{6}$ | 3.70 | 3.00 | 3.05 | 1.82 | 17.57% | 2.70 | 1.58 | 27.03% | 3.05 | 2.17 | 17.57% |
| | $\bar{6}$ | 45.70 | 36.06 | 6.60 | 4.10 | 85.56% | 6.40 | 3.3 | 86.00% | 5.70 | 2.17 | 87.53% |
| | $\widetilde{12}$ | 7.85 | 7.55 | 6.25 | 4.19 | 20.38% | 5.70 | 4.2 | 27.39% | 4.90 | 3.3 | 37.58% |
| | $\overline{12}$ | 25.85 | 19.77 | 9.05 | 2.76 | 64.99% | 13.05 | 5.32 | 49.52% | 9.00 | 6.64 | 65.18% |

## 4.3(a) grep 程序 NF 度量标准实验结果

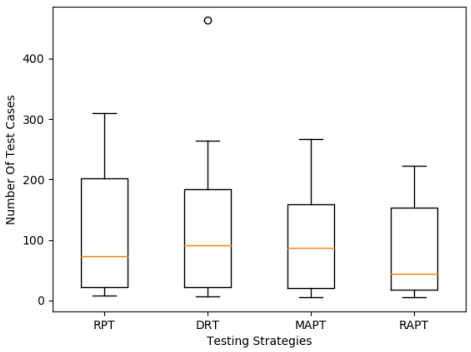
| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NF | SD_NF | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate |
| grepV1 | $\tilde{3}$ | 633.80 | 535.82 | 526.55 | 661.37 | 16.92% | 491.35 | 533.99 | 22.48% | 343.65 | 270.63 | 45.78% |
| | $\bar{3}$ | 429.60 | 373.72 | 412.85 | 471.55 | 3.90% | 448.70 | 344.39 | -4.45% | 375.40 | 313.63 | 12.62% |
| | $\tilde{9}$ | 401.55 | 286.02 | 349.80 | 285.71 | 12.89% | 320.50 | 336.94 | 20.18% | 315.15 | 383.04 | 21.52% |
| | $\bar{9}$ | 515.70 | 395.76 | 463.21 | 419.70 | 10.18% | 438.25 | 381.69 | 15.02% | 431.68 | 320.36 | 16.29% |
| | $\widehat{13}$ | 115.5 | 181.65 | 186.30 | 215.40 | -61.30% | 182.90 | 222.70 | -58.35% | 162.50 | 158.67 | -40.69% |
| | $\overline{13}$ | 522.05 | 466.9 | 427.00 | 260.66 | 18.21% | 494.55 | 522.46 | 5.27% | 492.25 | 413.38 | 5.71% |
| grepV2 | $\tilde{3}$ | 299.75 | 267.85 | 432.65 | 346.31 | -44.34% | 378.85 | 346.01 | -26.39% | 313.60 | 347.61 | -4.62% |
| | $\bar{3}$ | 342.30 | 406.57 | 276.05 | 321.14 | 19.35% | 269.85 | 268.09 | 21.17% | 206.30 | 191.94 | 39.73% |
| | $\tilde{9}$ | 6.85 | 7.48 | 6.40 | 8.78 | 6.57% | 6.15 | 6.48 | 10.22% | 3.40 | 4.16 | 50.36% |
| | $\bar{9}$ | 438.15 | 428.71 | 188.20 | 126.18 | 57.05% | 175.60 | 170.79 | 59.92% | 160.40 | 102.98 | 63.39% |
| | $\widehat{13}$ | 18.75 | 16.37 | 15.05 | 16.93 | 19.73% | 11.30 | 11.08 | 39.73% | 13.40 | 13.20 | 28.53% |
| | $\overline{13}$ | 170.35 | 188.04 | 115.25 | 91.20 | 32.35% | 164.20 | 151.76 | 3.61% | 102.10 | 77.29 | 40.06% |
| grepV3 | $\tilde{3}$ | 169.20 | 146.26 | 192.10 | 204.01 | -13.53% | 218.85 | 136.27 | -29.34% | 106.95 | 106.50 | 36.79% |
| | $\bar{3}$ | 130.20 | 118.86 | 119.80 | 111.16 | 7.99% | 103.10 | 70.09 | 20.81% | 83.30 | 79.10 | 36.02% |
| | $\tilde{9}$ | 102.40 | 77.71 | 151.80 | 158.06 | -48.24% | 149.65 | 140.16 | -46.14% | 145.55 | 99.16 | -42.14% |
| | $\bar{9}$ | 133.05 | 122.28 | 179.00 | 151.65 | -34.54% | 116.25 | 70.79 | 12.63% | 135.55 | 99.26 | -1.88% |
| | $\widehat{13}$ | 226.00 | 227.03 | 148.50 | 134.51 | 34.29% | 144.60 | 126.03 | 36.02% | 106.55 | 103.60 | 52.85% |
| | $\overline{13}$ | 214.70 | 174.46 | 218.85 | 136.27 | -1.93% | 123.15 | 119.64 | 42.64% | 205.05 | 227.47 | 4.49% |

## 4.3(b) make 程序 NF 度量标准实验结果

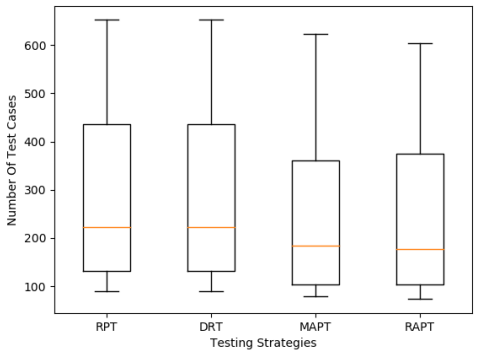
| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NF | SD_NF | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate |
| makeV1 | $\tilde{4}$ | 201.20 | 225.46 | 187.25 | 212.33 | 6.93% | 177.80 | 214.98 | 11.63% | 186.00 | 154.88 | 7.55% |
| | $\bar{4}$ | 276.20 | 306.01 | 184.15 | 182.68 | 33.33% | 135.95 | 192.69 | 50.78% | 181.75 | 118.80 | 34.20% |
| | $\tilde{6}$ | 229.10 | 335.26 | 285.00 | 228.91 | -24.40% | 263.70 | 206.50 | -15.10% | 245.10 | 518.65 | -6.98% |
| | $\bar{6}$ | 320.20 | 296.09 | 203.15 | 168.65 | 36.56% | 196.65 | 135.35 | 38.59% | 183.25 | 185.36 | 42.77% |
| | $\widehat{12}$ | 405.10 | 351.37 | 298.05 | 325.79 | 26.43% | 260.75 | 373.50 | 35.63% | 177.50 | 223.28 | 56.18% |
| | $\overline{12}$ | 293.90 | 156.18 | 331.50 | 319.12 | -12.79% | 250.50 | 244.02 | 14.77% | 167.85 | 196.71 | 42.89% |

## 4.3(c) gzip 程序 NF 度量标准实验结果

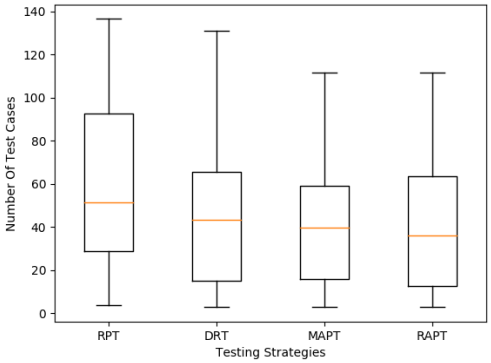
| Subject | initial profile | RPT | | DRT | | | MAPT | | | RAPT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NF | SD_NF | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate | NF | SD_NF | improvement rate |
| gzipV1 | $\tilde{4}$ | 73.05 | 64.32 | 82.80 | 85.64 | -13.35% | 62.90 | 61.18 | 13.89% | 34.80 | 29.30 | 52.36% |
| | $\bar{4}$ | 162.75 | 184.63 | 91.20 | 90.36 | 43.96% | 68.65 | 43.22 | 57.82% | 72.05 | 58.09 | 55.73% |
| | $\tilde{6}$ | 110.50 | 78.39 | 130.30 | 124.91 | -17.92% | 107.60 | 100.15 | 2.62% | 131.05 | 202.52 | -18.60% |
| | $\bar{6}$ | 262.35 | 189.71 | 126.35 | 87.82 | 51.84% | 170.90 | 142.43 | 34.86% | 166.60 | 238.03 | 36.50% |
| | $\widetilde{12}$ | 270.15 | 208.58 | 212.70 | 166.38 | 21.27% | 178.35 | 198.04 | 33.98% | 176.10 | 249.36 | 34.81% |
| | $\overline{12}$ | 193.40 | 281.05 | 201.95 | 224.23 | -4.42% | 183.30 | 181.48 | 5.22% | 180.70 | 208.15 | 6.57% |
| gzipV4 | $\tilde{4}$ | 71.95 | 66.61 | 61.00 | 56.41 | 15.22% | 54.45 | 43.30 | 24.32% | 41.05 | 39.54 | 42.95% |
| | $\bar{4}$ | 253.35 | 218.50 | 72.90 | 53.50 | 71.23% | 63.00 | 36.91 | 75.13% | 67.05 | 56.50 | 73.53% |
| | $\tilde{6}$ | 62.10 | 61.61 | 61.10 | 56.79 | 1.61% | 54.00 | 61.30 | 13.04% | 44.10 | 61.25 | 28.99% |
| | $\bar{6}$ | 177.50 | 190.54 | 107.55 | 64.03 | 39.41% | 112.90 | 84.75 | 36.39% | 71.85 | 67.60 | 59.52% |
| | $\widetilde{12}$ | 207.80 | 141.58 | 172.15 | 129.65 | 17.16% | 170.50 | 137.86 | 17.95% | 160.85 | 169.96 | 22.59% |
| | $\overline{12}$ | 281.95 | 251.63 | 174.65 | 201.61 | 38.06% | 155.75 | 154.13 | 44.76% | 148.65 | 134.98 | 47.28% |
| gzipV5 | $\tilde{4}$ | 10.05 | 11.11 | 8.55 | 5.68 | 14.93% | 8.40 | 6.59 | 16.42% | 6.35 | 7.06 | 36.82% |
| | $\bar{4}$ | 53.30 | 48.54 | 12.40 | 6.56 | 76.74% | 23.65 | 24.64 | 55.63% | 11.50 | 10.96 | 78.42% |
| | $\tilde{6}$ | 5.45 | 3.78 | 4.80 | 4.62 | 11.93% | 4.60 | 6.33 | 15.60% | 5.25 | 3.46 | 3.67% |
| | $\bar{6}$ | 71.50 | 61.55 | 6.75 | 5.30 | 90.56% | 6.90 | 3.37 | 90.35% | 6.55 | 3.02 | 90.84% |
| | $\widetilde{12}$ | 10.85 | 12.19 | 10.20 | 6.73 | 5.99% | 10.05 | 10.15 | 7.37% | 7.70 | 4.74 | 29.03% |
| | $\overline{12}$ | 111.10 | 91.40 | 12.45 | 6.85 | 88.79% | 17.60 | 10.35 | 84.16% | 10.50 | 5.12 | 90.55% |




(a) grep 实验

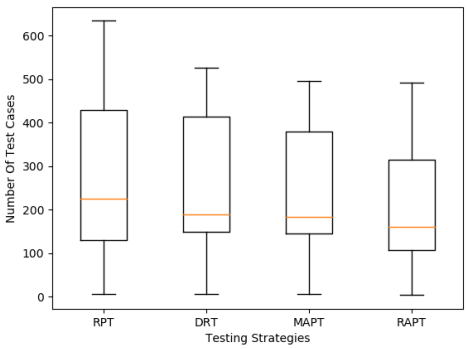



(b) make 实验

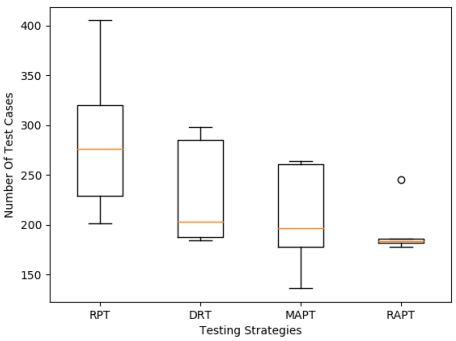



(c) gzip 实验

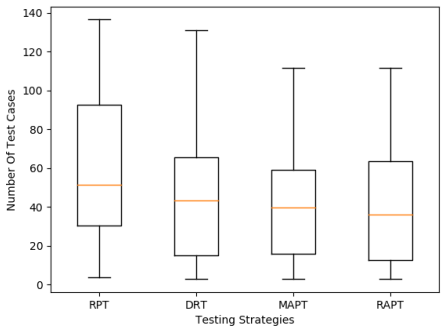
图 4.1 不同的实验对象 F 度量标准 RPT、DRT、MAPT、RAPT 策略的测试结果




(b) grep 实验




(a) make 实验




(c) gzip 实验

图 4.2 不同的实验对象 NF 度量标准 RPT、DRT、MAPT、RAPT 策略的测试结果

# 5 数据分析和讨论

## 5.1 数据分析

表 4.2 和表 4.3 给出了 3 个实验对象不同版本下的 RPT、DRT、MAPT、RAPT 测试策略在 F 度量指标以及 NF 度量指标下的测试结果$\bar{F}$和$\overline{NF}$标准差SD_F和SD_NF。DRT、MAPT、RAPT 相对于 RPT 策略的提升率被计算并且展示在该表中。

从表 4.2 和表 4.3 中我们可以观察到如下信息：

1. 与 RPT 相比，DRT 策略在大多数情况下揭示软件中前两个故障表现地更好。特别地，DRT 策略最高的提升率为 90.56%，最低的为-61.30%。提升率为负值表明此时 DRT 策略的表现不如 RPT 策略，在揭示第一个错误时比 RPT 策略多用了 61.30%的测试用例。提升率最高的情况出现在 gzipV5 实验对象初始条件为$\bar{6}$，此时能够揭示故障的测试用例集中在一个相对较大(测试用例数目较多)的分区之中，因此在不均等条件下该分区具有相对较高的概率被选中。另一方面，能够揭示故障的测试用例

集中在这一分区之中使得该分区的失效率也比较大。综合这两方面因素使得 DRT 策略在该实验对象下表现最突出。另外在标准差方面，DRT 策略在大多数情况下比 RPT 低，说明在测试过程中 DRT 比 RPT 更加稳定。但是表中仍然可以清晰地看到负值，说明此时 DRT 策略在重复测试地过程中测试结果波动较大，但是并不能说明 DRT 策略不如 RPT。事实上，本文在比较不同测试地效率时重点考虑的是$\overline{F}$和$\overline{NF}$。

2. 与 RPT、DRT 相比，MAPT 策略揭示软件前两个故障表现的更好。特别地，MAPT 策略最高地提升率为 90.35%，最低的为-58.35%。从图 4.1 中可以看出 MAPT 策略在几乎所有的情况下表现比 RPT 策略更好。并且从表 4.2 和表 4.3 中可以看出 MAPT 策略在大多数情况下比 DRT 表现地更好。表 4.2 中 gzipV5 实验在初始配置为$\overline{4}$以及$\overline{12}$时，MAPT 相对 DRT 策略表现明显不好。原因可能如下：从表 4.2(b)中可以看出在以上两种初始配置下，DRT 策略只需要 10 个左右的测试用例就能揭示软件中的故障，由于所需要的测试用例数目很小 MAPT 策略可能还有没来的及发挥它的优势；另一方面从表 4.2(b)中可以看出 MAPT 策略在这两种情况下的方差比 DRT 策略的方差大，说明在重复实验的过程中存在一些情况：MAPT 策略需要稍微多一点的测试用例揭示软件中的故障，使得整体均值比 DRT 大一些。但是在绝大多数情况下，MAPT 策略在揭示软件中的第一个故障时表现的更好。
3. 与 RPT、DRT 相比，RAPT 策略揭示软件前两个故障表现的更好。特别地，RAPT 策略最高地提升率为 90.84%，最低的为-42.14%。从表 4.2 和表 4.33 中可以看出 RAPT 策略在几乎所有的情况下比 RPT 策略表现的更好，并且 RAPT 策略在大多数情况下比 DRT 表现地更好
4. 表 4.2 和表 4.3 可以看出，均等与不均等的初始概率分布作为初始测试剖面对不同测试策略的效率影响没有明显的规律。这是由于测试剖面反应了不同分区的故障检测能力，但是实际的测试过程中并不能事先知道分区的故障检测能力。并且根据分区内测试用力数目多，该分区具有更高的故障检测能力不一定是正确的，很有可能造成不好的结果，例如表 4.2 中 gzipv1 实验。也可能带来好的结果例如表 4.2 中 gzipv2。但是并不能确定在一定的情况下哪一种设置测试剖面的方式更好。测试工程师可以根据以往的测试经验设置测试剖面。但是如果没有以往的测试信息保守地可以将均等初始概率分布作为初始测试剖面。

从上面的数据分析可以得到如下结论：

1) DRT 策略在大多数情况下揭示前两个故障比 RPT 策略表现地更好。

2) MAPT 策略以及 RAPT 策略揭示前两个故障地能力高于 RPT 策略以及 DRT 策略。

## 5.2 讨论

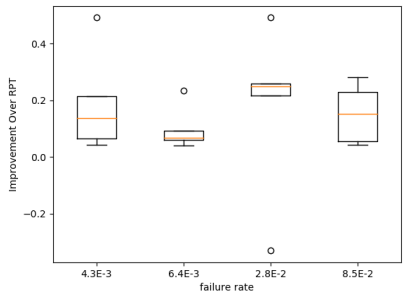
基于理论分析与实验结果，提出几个问题并进行讨论。

1. 待测软件输入域的失效率对不同测试策略测试效率的影响。

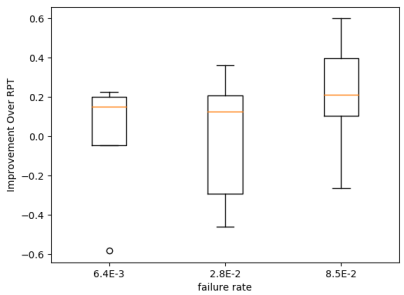
本文植入待测软件中的故障都是相对难检测的，各个实验的输入域的失效率如表 5.1。各个实验对象在不同的失效率下 MAPT、RAPT 策略相对于 RPT 策略的效率提升率得到图 5.1。

| Subject | versions | overrall failure rate |
|---|---|---|
| grep | V1 | 6.4E-3 |
| | V2 | 8.5E-2 |
| | V3 | 2.8E-2 |
| | V4 | 4.3E-3 |
| gzip | V1 | 9.3E-3 |
| | V2 | 6.1E-2 |
| | V4 | 4.7E-3 |
| | V5 | 2.3E-2 |
| make | V1 | 1.0E-2 |
| | V2 | 2.5E-3 |

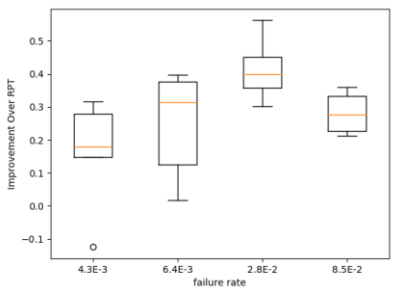
表 5.1 待测程序失效率信息



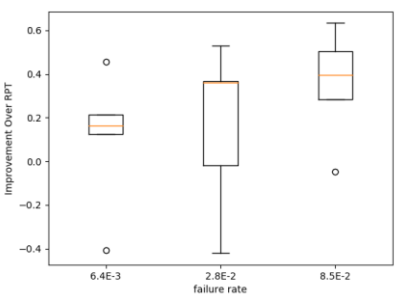

(a) grep 实验 MAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率



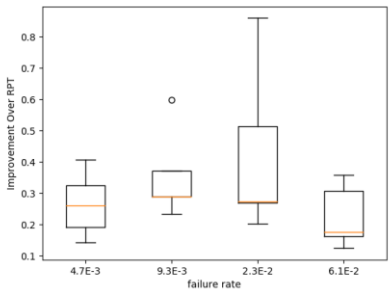

(b) grep 实验 MAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率



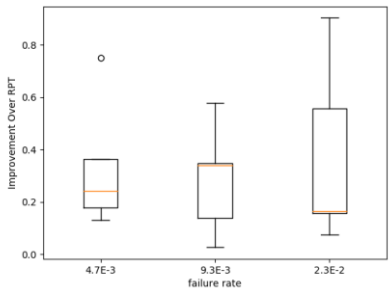

(c) grep 实验 RAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率



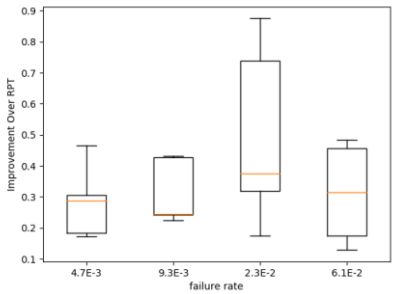

(d) grep 实验 RAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率



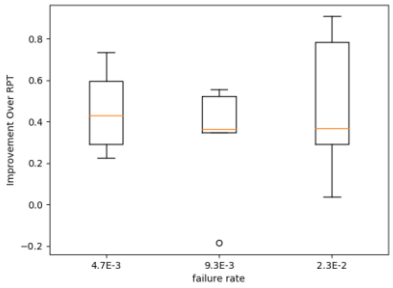

(e) gzip 实验 MAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率



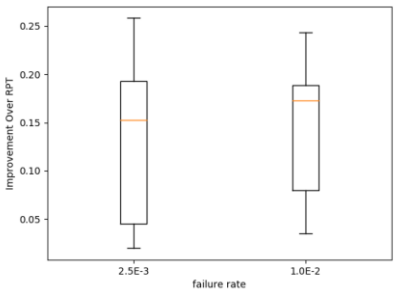

(f) gzip 实验 MAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率



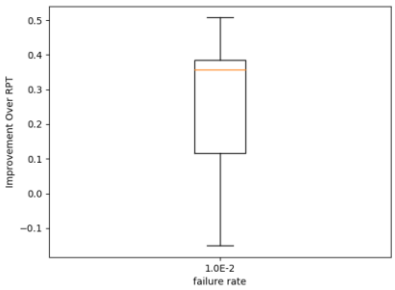

(g) gzip 实验 RAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率



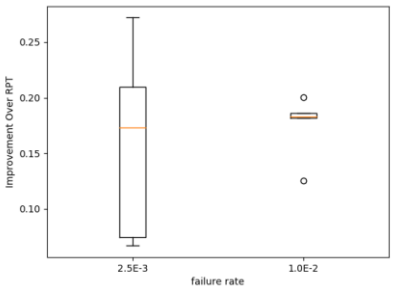

(h) gzip 实验 RAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率



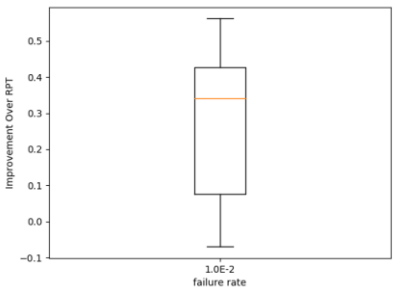

(i) make 实验 MAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率




(j) make 实验 MAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率




(k) make 实验 RAPT 在不同失效率下 F 度量标准相对 RPT 策略的提升率




(l) make 实验 RAPT 在不同失效率下 NF 度量标准相对 RPT 策略的提升率

图 5.1 不同实验对象 MAPT、RAPT 策略在不同失效率下 F、NF 度量标准相对于 RPT 的效率提升率

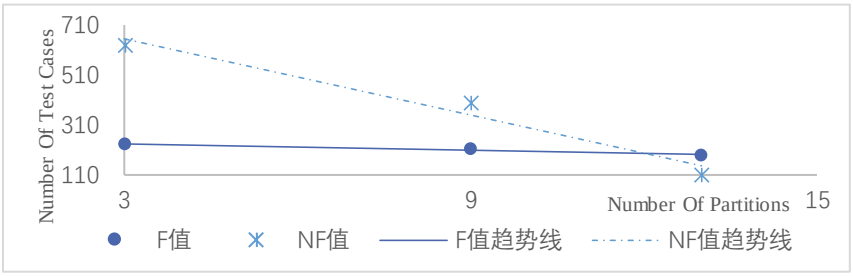
根据图 5.1 所示当软件中的失效率在[0.0025, 0.0851]之间时，本文的实验并不是仿真实验因此不能使实验对象的失效率在[0.0025, 0.0851]之间，以恒定的增长值依次出现。从途中可以看出当软件的失效率在[0.0025, 0.0851]之间时，MAPT 与 RAPT 策略比 RPT 策略均有明显的提高。

2. 分区数目对不同测试策略测试效率的影响。

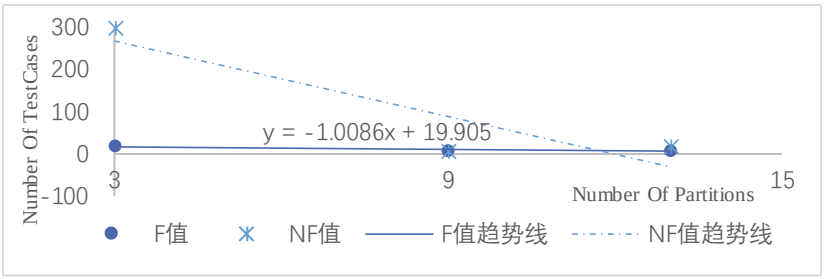
上文讨论了均等与不均等地初始概率分布作为初始测试剖面对测试策略效率地影响是不确定的。因此本节只讨论在均等的初始概率分布作为初始测试剖面时，分区数目对测试策略的影响。

问题 1 讨论了不同失效率宜采用的测试策略问题，本节讨论选取测试策略之后，分区数目的选择。在实际测试中，很难讲哪一种分区策略能够提高测试效率，并且同一种分区策略可以得到不同的分区方式。例如本文在第四章提到的以分区的粗细程度得到每一个待测程序的 3 种分区方式：粗粒度、中等粒度、细粒度。并且分区的粒度越细则分区的数目就越多。图 5.2 展示了 RPT 策略在某一个待测程序中，不同分区数目的测试结果。图中横轴代表分区数目，纵轴代表揭示软件中的第一个故障所需要的测试用例的平均数。本文用 RPT 策略的实验结果研究分区数目对测试效率的影响原因是：DRT、MAPT、RAPT 相对于 RPT 具有更高的故障检测能力，因此 DRT、MAPT、RAPT 策略可能需要相对于 RPT 较少的测试用例揭示故障但是整体趋势应当和 RPT 策略相似。

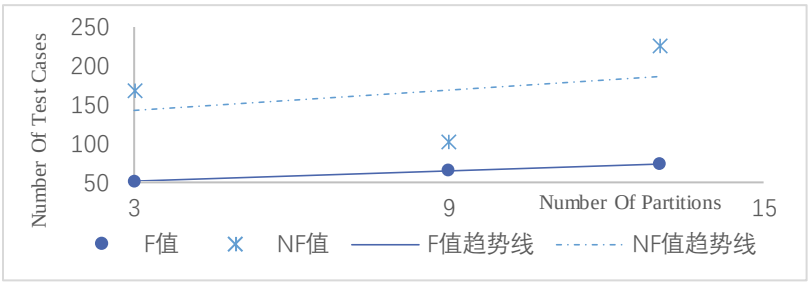
从表 5.1 可以看出 grepV2 以及 gzipV2 的失效率分别为 8.5E-2 和 6.1E-2 明显比其它实验对象的失效率要大。从图 5.2(b)、(f)可以明显看到 grepV2 实验 F 和 NF 度量标准的大小与分区数目成明显的负相关。原因可能是：失效率比较大，说明可以揭示故障的分区以及测试用例数目比较多，当分区的粒度大时，分区内的测试用例数目相对较少，使得具有故障检测能力的分区的失效率比较大。另一方面，具有故障检测能力的分区数目比较多使得选中具有故障检测能力的分区就相对容易。其它实验对象 F 和 NF 度量标准的大小与分区数目成明显的正相关。原因可能是：首先失效率比较低意味着能够揭示故障的测试用例数目比较少并且具有故障揭示能力的分区数目可能也比较少。当分区数目增多时可能使得具有故障检测能力的分区失效率增大但同时也使得选中具有故障检测能力的分区更加困难。由此可能导致揭示故障所需要的测试用例数目随着分区数目增多而增加。图 5.2(a)可以看出该实验对象的从测试用例集的失效率比较小为 6.4E-3 但是揭示故障所需要的测试用例数目与分区数量呈现明显的负相关。原因可能是：该实验对象的三种分区方式下均具有较多的具有故障检测能力的分区，因此随着分区数目的增加具有故障检测能力的分区的失效率变大，因此选中能够揭示故障的测试用例也越来越容易。



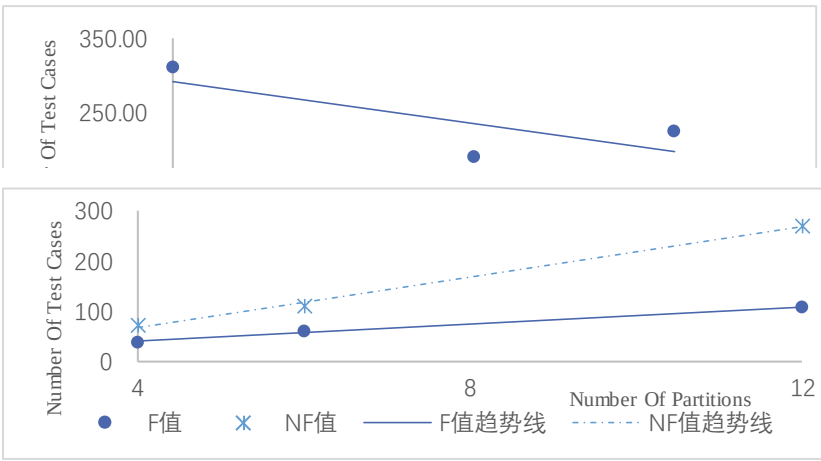

(a) grepV1 均等初始概率分布



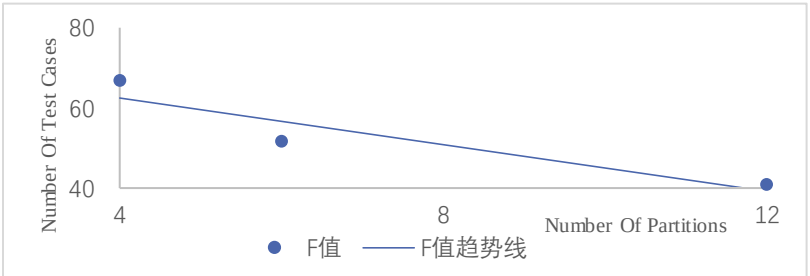

(b) grepV2 均等初始概率分布



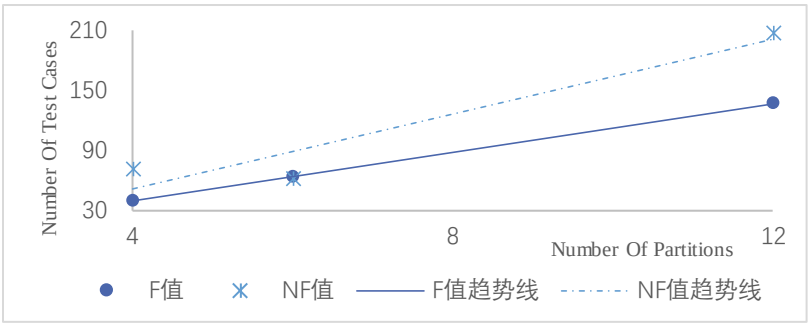

(c) grepV3 均等初始概率分布



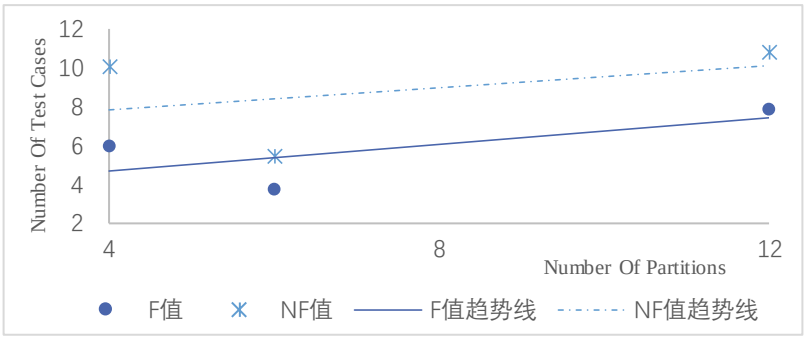

(e) gzipV1 均等初始概率分布



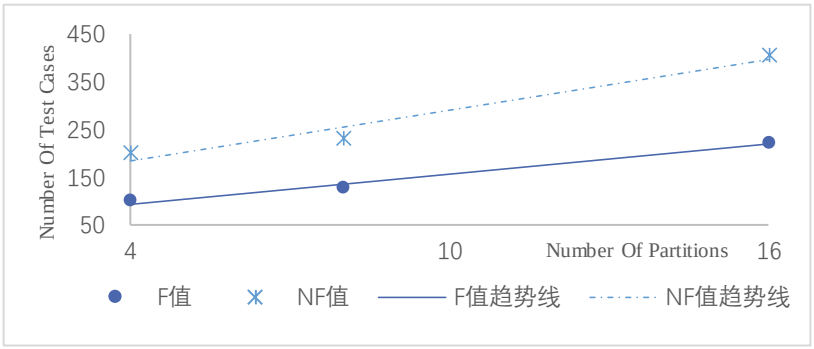

(f) gzipV2 均等初始概率分布



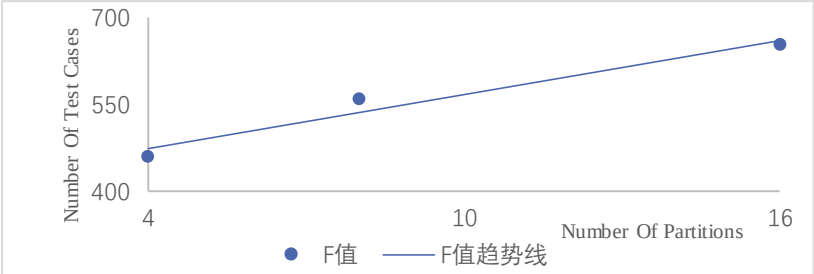

(g) gzipV4 均等初始概率分布




(h) gzipV5 均等初始概率分布




(i) makeV1 均等初始概率分布




(j) makeV2 均等初始概率分布

图 5.2 RPT 策略在不同实验对象下不同分区数目的测试结果

1. Computational overhead

从上面的讨论可以看出 DRT、MAPT、RAPT 相对于 RPT 策略的测试效率有明显的提高，但是也带来额外的计算开销。表 5.2 记录了 makeV1 实验对象 16 个

分区数目并且以均等的初始概率分布作为初始测试剖面下各个测试策略的计算开销。由于揭示软件中的故障所需要的测试用例的数目相对较多，因此不同测试策略的计算开销可以清楚的看到。

表 5.2 COMPUTATIONAL OVERHEAD(makeV1)

| | RPT | DRT | MAPT | RAPT |
|---|---|---|---|---|
| $\bar{F}$ | 198.2 | 181.55 | 175.25 | 158.2 |
| F_Total(ms) | 11542.75 | 11192.0 | 11569.1 | 9756.3 |
| $\overline{NF}$ | 423.35 | 297.75 | 267.9 | 280.6 |
| NF_Total(ms) | 23967.5 | 17802.0 | 17287.2 | 16455.55 |

这个实验是在虚拟机上 Ubuntu 11.04 64-bit 的操作系统，该系统分配了 2 个处理器，2GB 的内存。测试脚本用的是 Java 和 bash shell。正如表 5.2 展示的三种测试策略在 F 以及 NF 度量指标下所需要的时间均比 RPT 策略的少，由此可以看出 DRT、MAPT、RAPT 相对于 RPT 的额外计算开销可以通过减少揭示故障的测试用例数目弥补。MAPT 以及 RAPT 策略相对于 DRT 策略的计算开销相差不大。

# 6 结论和将来的工作

动态随机测试是一个旨在利用历史的测试信息动态改变测试剖面的测试策略。DRT 策略的主要优点是测试剖面不断变化，使得较高失效率的分区具有更高的被选择概率。但是 DRT 策略的测试效率受分区数目、初始剖面这些外部因素的影响。同时 DRT 策略的测试效率也受内部机制的影响：该策略根据某一个分区的执行结果调整所有分区被选择的概率并且所有的分区调整概率的幅度都相同。本文结合 Markov 链的状态转移矩阵提出了 MAPT 策略缓解了 DRT 策略的内部机制的不恰当问题。由于传统的 DRT 策略的参数取值普遍很小，并且分区被选择的概率容易受其它分区测试结果的影响使得找出具有较高故障检测能力的分区的速度较慢。本文提出基于奖惩机制的 RAPT 策略缓解这一问题。针对 DRT 策略的两个外部影响因素本文为每一个实验设置了不同数目的分区，并且为每一种分区方式设置均等的初始概率分布和不均等的初始概率分布作为初始剖面。通过对 3 个真实的程序的不同版本进行测试，实验结果表明 MAPT 策略以及 RAPT 策略比 DRT 以及 RPT 具有更高的故障检测能力。当失效率比较高时简单的测试策略 RPT 或许是一个更加合理的选择。当失效率较低时意味着需要很多的测试用例揭示软件中的故障，此时用更少的测试用例揭示软件中的故障无疑要比增加的计算代价划算。因此失效率低时采用 MAPT、RAPT 策略较为合理。本文通过实验发现：当待测软件的故障不容易揭示时，宜采用粗粒度的分区方式；当待测

软件的故障容易揭示时，宜采用较多数目的分区方式。

但是 MAPT 策略中参数$\tau,\gamma$应当满足$\gamma > \tau$，因为在实际情况中输入造成的故障要比输入没有造成故障少。RAPT 策略中惩罚上限的设置不同的实验可能取值不同，对策略的测试效率也有影响。将来的重点工作是研究 MAPT 策略中的参数以及 RAPT 策略中的惩罚上限进一步提高 MAPT、RAPT 策略的测试效率。

# 7 参考文献


[1] R. Hamlet, “Random Testing,” Encyclopedia of Software Engineering, J. Marciniak(ed), Wiled, pp. 970-978, 1994.

[2] W. J. Gutjahr, “Partition Testing vs Random Testing: The Influence of Uncertainty,” IEEE Transactions on Software Engineering, vol. 25, issue 5 pp: 661-674, Setember/October 1999.

[3] R. Hamlet and R. Taylor, “Partition Testing Does Not Inspire Confidence,” IEEE Transactions on Software Engineering, vol. 16, no. 12, pp. 1, 402-1, 411, December 1990.

[4] T. Y. Chen and Y. T. Yu, “Amore general sufficient condition for partition testing to be better than testing,” Information Processing Letters, vol.57, issue 3, pp. 145-149, February 1996.

[5] K. Y. Cai, B. Gu, H. Hu, and Y. C. Li, “Adaptive Software Testing with Fixed-Memory Feedback,” Jounal of Systems and Software, vol. 80, issue 8, pp. 1328-1348, August 2007.

[6] K. Y. Cai, T. Jing, and C. G. Bai, “Partition Testing with Dynamic Partitionint,” in Proceedings of the 29th COMPSAC, Edinburgh, Scotland, July 2005, pp. 113-116.

[7] A. G. koru, K. E. Emam, D. S. Zhang, H. Liu, D. Mathew, “Theory of relative defect proneness,” Encyclopedia of Software Engineering, vol. 13, no. 5, pp. 473-498, 2008.

[8] P. E. Ammannn and J. C. Knight, “data diversity: an approach to software fault tolerance,” IEEE Transactions on Computers, vol. 37, no. 4, pp. 418-425, April 1988.

[9] G. B. Finelli, “NASA Software Failure Characterization Experiments,” Reliability Engineering and System Safety, vol. 32, pp. 155-169, 1991.

[10] K. Y. Cai, “Optimal software testing and adaptive software testing in the context of software cybernetics,” Information and Software Technology, vol. 44, pp. 841-855, November 2002.

[11] K. Y. Cai, H. Hu, C. H. Jiang, and F. Ye, “Random testing with dynamically updated test profile,” in Proceedings of the 20th ISSRE 2009, Fast Abstract 198, Mysuru, Karnataka, India, November 2009.

[12] E. J. Weyuker and B. Jeng, “Analyzing partition testing strategies,” IEEE Transactions on Software Engineering, 17(7): 703-711, July 1991.

[13] Junpeng Lv, Hai Hu, and Kai Yuan Cai, “A Sufficient Condition for Parameters Estimation

in Dynamic Random Testing," International Computer Software and Application Conference, 2011 IEEE 35th Annual, Munich, 2011, pp. 19-24.

[14] Z. J. Yang, B. B. Yin, J. P. Lv, K. Y. Cai, S. S. Yau, and J. Yu, "Dynamic Random Testing with Parameter Adjustment," COMPSACW, 2014 IEEE 38th Annual, 2014, pp. 37-42.

[15] T. J. Ostrand, M. J. Balcer, "The Category-Partition Method for Specifying and Generating Functional Tests," Communications of the ACM, vol. 331, pp. 676-686, June 1988.