Undergo a Major Revision. 14-Feb-2018.  
1. “Summary of Changes”. 2. Response to Comments. 3. Identify a teaser image 4. Revision.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
Editor Comments  
  
The referee believes that some of the contributions are not adequately supported by the empirical evidence, and therefore questions whether the paper overall has enough of a contribution to form the basis of a publication in the journal.  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Reviewers' Comments  
  
Reviewer: 1

Comment 1:

However, can the authors clarify whether and why the feedback checking mechanism is something inherent to their suggested approach, or rather orthogonal that can be embedded into SCT and FDA-CIT as well? Especially considering the conclusion for Q5 that ICT handles better the cases where non-determinism and non-safe values exist thanks to this feedback checking process.

**Response: Alright. We have added one paragraph in Section 4.2.1 (Page 9, the red part) to clarify this. Our main point is that this feedback checking mechanism can also be embedded into SCT. However, there are two drawbacks that can limit the improvement of this mechanism. First, the fix of wrongly identified MFS cannot be used in other test cases. Second, it costs SCT more to embed feedback checking mechanism.**

**With respect to FDA-CIT, our feedback checking mechanism cannot be directly applied on it. This is mainly because the MFS identification process of FDA-CIT is much different from that of ours, which is a post-analysis classification tree method. It does not need any additional test cases, while our approach needs additional test cases to refine the accurateness of MFS identification. Another problem to apply feedback checking mechanism on FDA-CIT is that it may cause the over-fitting issues of the classification tree method of FDA-CIT.**

Comment 2:

I feel that the statement "Traditional covering array usually offer an inadequate testing due to Masking effects" (3.3) needs to be refined. This is true only when considering a single execution of the test set. In practice, it is often the case that the set is rerun until all test cases pass. A passing regression is even a commonly defined gating condition in Cloud development. [5] mentions fix and rerun as a viable option to handle masking effects, though of course it has its limitations - it would be better to progress with testing and cover as much as possible while waiting for the fixes to be available, which is what ICT and FDA-CIT offer.

**Response: We agree. We have refined this sentence according to your comments. The changes can be found in Section 3.3 (red part of paragraph 1 and 2). Specifically, we have emphasized the point that “Running traditional covering array at one time is inadequate, and testers usually try to repeatedly run the covering array in practice to achieve a relatively adequate testing”.**

Comment 3:

Finally, I do not understand the reasoning for the choice on how to handle constraints. The authors indicate they choose the dynamic method since "it can directly \*be\* applied into our framework". However it seems to me that both methods can be applied to the framework. If you already have a mechanism for handling forbidden tuples for previously discovered MFS, does it really matter if you provide forbidden tuples during initialization or on-the-fly? Where it seems to matter is in the high performance penalty you would pay by using the dynamic method to discover the minimal forbidden tuples one by one through invalid test generation, when in typical real-world models hundreds or even thousands of such tuples exist (in many cases more than the number of failures). In the static approach, the user needs to pre-define constraints, but usually these are known, and each constraint typically concisely summarizes a large set of forbidden tuples. You could also use a combined approach, where the user specifies known constraints even if they are incomplete in order to reduce their appearance during execution and achieve performance savings.

**Response: It is a very good question, and we agree with you that both methods can be applied to the framework. Your suggestion that combining both static and dynamic ways of handling constraints is very useful. We believe that, in practice, it is wise to first give initial settings of constraints (especially when there are too many constraints) such that the generation approach can efficiently avoid invalid test cases, otherwise we need to identify these constraints one by one. The dynamic way of identifying constraints and then avoiding them is also useful, because we may not always get an accurate input model at the first time and some constraints may be ignored at some time.**

**According to your comments, we have rephrased the sentences for describing why we chose the dynamic way of handling constraints in our empirical studies (See Section 4.2.2, paragraph 1, the red part). Our main point is that there are two reasons for this choice. First, there are not many constraints in our empirical study such that the dynamic way of identifying them and forbidding them will not affect the efficiency too much. Second, the dynamic process of handling constraints is similar to the way that we identify the MFS, so our framework does not need to be modified a lot for handling constraints.**

Reviewer: 2  
  
Public Comments (these will be made available to the author)  
The authors addressed my comments well and I am happy to suggest the acceptance of the paper.

**Thanks a lot**!  
  
Reviewer: 3  
  
  
Comment 1:

One contribution they claim to have is the feedback checking mechanism given in Algorithm 2, which after identifying an MFS, tests a predefined number of test cases to check to see if a condition containing MFS really fails or not. When they are choosing these additional test cases they try to pick the ones that are the most different than the original failing test case, which seems to be an idea already studied in another paper. However, in the experiments no data is provided to show how much the proposed approach gained from this functionality. For example, what is the performance of the proposed approach with and without this functionality?? Therefore, it is not clear how valuable this contribution is.

**Response: According to this comment, we conducted an experiment to evaluate the approach without this feedback checking mechanism. Specifically, , we created a mutation version of ict by removing the feedback checking mechanism from the original ict approach, and then applied this approach on testing the software subjects and identifying the MFS. The results of our experiment show that, without feed-back checking mechanism, the number of test cases generated by ict reduced, but the quality of MFS identification and tested-t-way coverage decreased significantly. It indicates that the additional test cases generated in feedback checking mechanism is worthwhile, and it is beneficial to adopt the feedback checking mechanism in the CT process in order to obtain a better MFS identification result and a higher tested-t-way coverage.**

//实验 不确定错误  
One contribution they claim to have is the experiments they conducted measuring the performance of the proposed approach on non-deterministic failures. However, they did this using a quite questionable set of synthetic data where there was one deterministic and one non-deterministic failure. What was the point having the deterministic failure in the experiments. Didn’t it greatly increase the F-measures reported in the paper. Even so, the F-measures were significantly dropped and as noted in the paper between the non-determinism levels of 0.4 and 0.8 (where the failures manifested themselves between 40% and 80% of the time) it was about and below 0.4. Therefore, it doesn’t really make sense the performance of ICT (the proposed approach) to other approaches because an F-measure of 0.4 is not acceptable. When this coupled with the fact that had more non-deterministic MFSes were used, the F-measures would have been even smaller, the statement (page 20, 2nd column, lines 55-58) “this result indicates that act can handle the non-deterministic failures properly” becomes simply not acceptable.

使用了非常有疑问的人工数据。 为什么有一个 确定的 和一个非确定的。 当non-deterministic 程度到达0.4和0.8的时候，还有很大的降低 （低于0.4）。 不一定好于其他。因为0.4不可接受。 因此，这个方法比其他方法好，能处理non-deterministic问题 这个结果是不能接受的。

Response: We have changed to , we found our results performed even better than the case. One explanation is that there is a voice.

这个问题我们分两个：

第一我们同意，我们改变了实验设计，抹去了那个deterministic的故障。我们发现我们的方法反而更好。

我们不同意。当我们的方法已经比其他方法都好的时候，没有

//对比实验 安全值  
Another claimed contribution is the experiments which were performed to evaluate ICT on models with no safe-values. First, these experiments were also carried out using a quite questionable set of synthetic data where for each value of a parameter was made to appear in an MFS. However, no information was provided about the cardinality of these manually generated MFSes. How many unique parameters each had?? This is important because if they had large cardinalities they were highly unlikely to be hit during testing.  Another question is how many times these MFSes actually caused failures during the MFS identification times?

使用了非常有疑问的人工数据。 每一个value都是一个MFS中的一部分。 但是不知道这些MFS的 大小， 每个参数有多少个独有的参数， 这个很重要，因为如果很大，那么不容易击中。 另一个问题是，这些MFS导致了多少错误在故障定位期间。

重新设计实验， 已 safe value 的cardinate为例。

Response: Information is provided. 包括大小，个数，引发故障个数。

So, multiple faults (defects) are not actually handled by the proposed approach. That is, the multiple MFSes used in the experiments were assumed to be caused by exactly the same defect. So, the system under test is assumed to have exactly one defect, which is indeed also evident in Table 7. First, this significantly hinders the applicability of the proposed approach as this is a quite strong of an assumption to have for a testing approach. Second, the effect of multiple faults/defects should have been studied. Third, the comparisons made against FDA-CIT is not fair as FDA-CIT seems to design to work with not only multiple defects but also with multiple defect-test case pairs.

**Response: As suggested by the reviewer, we have added the experiments with multiple defeats (See Section 5.6, blue part). Specifically, we considered five different software subjects with multiple defeats, and then conducted the three approaches, i.e., ict, sct, and fda-cit, on testing them and identifying the MFS. We found that most results matched pretty well with the results obtained from the experiments of single defeat. Particularly, ict performed better at MFS identification and reducing of number of test cases that containing multiple MFS, and fda-cit generated the least number of test cases and covered the most number of tested-t-way schemas. When compared with ict and fda-cit, approach sct was normally in between.**

I change the example: we have one defect causing multiple  
MFSes:  
abs  
000  
100  
But this time a is a ternary option and b ,c, and d are binary options. Since not all the values of a parameter will be exercised by ICT, starting with (0, 0, 0, 0), we would have  
abcd  
0000 F  
1000 F  
0100 P  
0010 P  
0001 F  
(correct me if I am wrong) which will mark (b=0, and c=0) as an MFS and it would be wrong! Note that this is different than the discussion about why finding the minimum MFS is expensive. This is simply the case where the algorithm wrongly identifies a combination as MFS. Therefore, the authors, I guess, could have come up something with an “approximate MFS - AMFS” definition and rather than stating that ICT finds MFS in many places in the text, which is definitely confusing and not seem to be right, as shown by the counter example above, could have stated that ICT finds AMFS.

Response: Unfortunately, this refined example you provided is still wrong. There are two mistakes you make in this example.

First, the safe value is not satisfied here. OFOT can identify the correct MFS only because the safe value assumption. It seems that the reviewer is not familiar with .

Second, our ict can work under the condition that safe value is not satisfied. So the correct of this is worked as following:

//算法没有给出 一些具体的函数没有给出。  
Part of the confusion here is that although two algorithms were given regarding certain functionalities in ICT the main algorithm was not actually given, so it is not possible how exactly ICT proceeds.

Response: our example is shown. And we explicitly show the algorithm.

//约束 单故障的, 以及单错误的 讨论  
Regarding the response to Reviewer 3’s comment #16: Authors states that “we explicitly show the assumption used in our study (See Section 2.2).” However, in Section 2.2. the only approach, the assumptions of which are explicitly states, is the OFOT approach. They don’t actually talk about the assumptions of ICT. Because for example Assumption #2 states basically states that no multiple MFS, but ICT (or at least the rest of the paper) is all about multiple MFSes. Furthermore, the single defect assumption we have discussed above is not discussed anywhere in the paper.

Response： 单故障只是确保OFOT的正确性， ICT 作为一个framework能有什么假设。而且通过feed-back 已经解决了这个问题。其实多故障只会使得我们的算法不准确。

单个错误的，我们已经给出了一种解法。

//相关的是 处理 多个MFS  
A related concern is that although ICT seems to be mostly about multiple MFSes, it does not seem to offer a mechanism to address this issue if a given failing test case has multiple MFSes. It simply attempts to avoid generating test cases with multiple MFSes by opportunistically hoping to hit test cases with single MFS before it hits test cases with multiple test cases, so that the single MFSes can be ruled out to reduce the chance of hitting multiple MFSes. Since there is no guarantee for this, show ICT performs on test cases with multiple MFSes is still a valid question? What would be the F-measure obtained for these cases?

Response: 1. our results, especially feed-back checking can remove this condition. Our empirical study can significantly reduce the case.

2. We have simply attached the results for. Our method can account it, but the results is still good than feed-back can check the multiple and just discard the problem. (Just cost more test cases).

//这个需要说为什么ofot好于 fic  
Regarding the response to Reviewer 3’s comment #8: Correct me if I am wrong, in the original submission, ICT used FIC in the MFS identification phase and the reason was that FIC performed better than OFOT. In this submission, ICT uses OFOT, because now it performs better. This could happen but I believe the authors should share their insights with the reader.

Response: we have emphasized.

First, we use OFOT is all because to satisfy your requirement that all the MFS identification approach used in this paper should be unified for fair comparison. (instead of “it now performs better”). Second, since our paper does not use FIC we do believe there is no need to explain why now use OFOT is better to the reader. Third, for you, we would like to explain that.

我们的结果更加好主要源于我们的新的算法。而不是定位。既然我正篇文章都没提 fic， 那就没必要说了。

//F-measure 要考虑没有定位到的mfs

Section 5.6.2, paragraph 2: This paragraph basically indicates that if a failure caused by an MFS is not hit then that MFS is not included in the F-measures. Why not? Isn’t it a failure that the approach missed the MFS? Even if for some reason they need to be discarded, in the experiments statistics about the number of failures and number of unique failures found by each approach should be given. Because the number of MFSes hit, will definitely affect the F-measures computed, which should not have been the case. Because this would favor an approach that hits, for example, one MFS out of many and correctly localizes it over an approach that hits all the MFSes but failed to correctly localize all of them.

这是一个明显的错误！否则我们的算法的f-measure就是100%，在小于0.01的时候，但是没有，就是因为没有击中那个mfs没有激发错误。

Response: we do not computed the MFS with out. Which sentence?

Furthermore, due to the all the assumptions made in the paper and the emphasis on doing this in a dynamic manner is much better than doing it in a static manner, the approach should have been compared to static Error Locating Arrays (ELAs). This should not be a problem because for the experiments carried out in this paper an ELA will be a standard (t+d)-way array, which will guarantee the identification of the MFSes up to and including the cardinality of d. So, the authors can generate standard covering arrays of appropriate strengths and compare their sizes to the numbers of test cases required by their approach.

**Response: As suggested by the reviewer, we have added one section (Section 5.9) to show the results of ELA. We obtained the same conclusion as [1]. That is, although ELA can identify the MFS accurately, it needs to generate much more test cases than the dynamic approaches. Additionally, it also needs to know the number and degree of the MFS at first, which limits its application in practice.**

**[1]C. Yilmaz, E. Dumlu, M. Cohen, and A. Porter, “Reducing masking effects in combinatorial interaction testing: A feedback driven adaptive approach,” Software Engineering, IEEE Transactions on, vol. 40, no. 1, pp. 43–66, Jan 2014.**

// 提出约束是如何具体处理的？  
  
There is a section on constraints in the paper (Section 4.2.2), which simply talks about the constraints identified by ICT. However, it is still not known how ICT picks values for the parameters if no values could be chosen due to the known and/or discovered constraints. For example, ICT wants to test a=0, but a=0 is invalided by some constraints. Then what???

Response: This is not valid. Our is forbidden, this is the implicated constraints. This is impossible, That is not. Is unsolution, we will update the constriants.

具体处理加算法，然后给出这个问题的解

//下面的都是详细解释算法和例子

It is stated in the paper (Section 4.4, paragraph 2) that if Algorithm 2 returns false (i.e., if one passing test case is found containing the previously identified MFS), the original failing test case is retested. How does this work? No detail is given in the paper.

Response: Re-tested is .

We give the detail of this algorithm.

In Section 34. Table 5, how does the augmented SCT actually generate t14 and t19 given the failing test case t4 and previously identified MFS (-,0,-,-)?? No detail is given about this in the paper. Again, the algorithms for ICT and SCT should be given in the paper.

Response: Table 5, the SCT generated t14 and t19 is because.

Do the algorithms and the equations, at least the ones given in the paper, really enumerate and store all valid test cases (all possible combinations of parameter values) as they suggest they do, which obviously not scalable at all???

Response: No, we do not really enumerate and store all valid test cases. Our eqution is just one way to simply show .

Summary:

第一个结束（comments 和 文章修改都结束了）

第三个reviewer的优先级：

Feedback 触发了多少次， 平均每次改正多少MFS identification， 消耗多少测试用例， 改善多少coverage。 （绿色的要在一起），

Ɵ 1.前三个关于之前的实验贡献 （需要添加实验）：

不确定错误的实验设计（减少一个确定的），

安全值的实验设计（有多少个 4维，然后告诉有多少触发了？然后有多少解决了）。

3. 自己给出的例子，我要给他解释他的错误。

4． 然后再次详细解释我们的算法，多个地方。 （关联 10， 11， 12， 13）

5. 单MFS假设说明 ， 只要回答他

Ɵ 6. 多个MFS的说明， 以及实验的f-measure在这些情况下的 值 （实验给出）

7. 为什么 ofot 反而好于 fic 给出说明

Ɵ 8. F-measure 要考虑没有定位到的mfs (实验给出？) 这是一个错误

10. 约束到底怎么处理的 （给出详细描述）

11. re-test 是怎么处理的

12. 那个例子中的 t 是怎么生成的

13. 那个真的要存储所有的validation的吗。

F-measure 。不同情况下（多个在一起，不确定，安全值）要分别记录统计 （这个是错误的）

增加两个实验。

今晚之前： feedback, multiple mfs. Ok

搞 underdetermine 和 safety 快 ok