Regression Test Suite Reduction using an Hybrid Technique
Based on BCO And Genetic Algorithm

Bharti Suri, Isha Mangal & Varun Srivastava
University School of Information Technology, Guru Gobind Singh Indraprastha University,
Dwarka, Delhi-110075, India
E-mail : bhartisuri@gmail.com, ishamangal2006@yahoo.com, varun0621@gmail.com

Abstract - Regression testing is a maintenance activity that is performed to ensure the validity of modified software. The activity takes a lot of time to run the entire test suite and is very expensive. Thus it becomes a necessity to choose the minimum set of test cases with the ability to cover all the faults in minimum time. The paper presents a new test case reduction hybrid technique based on Genetic algorithms (GA) and bee colony optimisation (BCO). GA is an evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. BCO is a swarm intelligence algorithm. The proposed approach adopts the behavior of bees to solve the given problem. It proves to be optimistic approach which provides optimum results in minimum time.

Key words - Regression testing; genetic algorithm (GA); bee colony optimisation (BCO).

I. INTRODUCTION

Software maintenance is an activity which includes enhancements, error corrections, optimization and deletion of obsolete capabilities [1, 2]. Regression testing is performed in maintenance phase and is defined [3] as “the process of retesting the modified parts of the software and ensuring that no new errors have been introduced into previously tested code”. It is an expensive activity and consumes significant amount of effort and cost. The test suits are already available from previous stages of software development life cycle (SDLC). Regression testing does not involve rerunning the entire suite but selecting a subset of it [4] that can detect all the faults. There are various regression techniques: Retest all [1, 5], Regression test selection [6-16], and test case prioritization [8, 17-19] and hybrid approaches [1, 2, 7]. As the presented paper is based on test suite selection; the paper discusses it in detail.

Regression Test Case Selection (RTS):

“Retest all” is very expensive technique, so regression test selection is performed to reduce the cost. The technique, instead of rerunning the whole test suite, choses a part of test suite such that cost of selecting a part of test suite is less than the cost of running the tests. RTS techniques are broadly classified into three categories [3] as in Fig 1: Coverage techniques [11], Minimization techniques and Safe techniques [3].

Fig. 1: Regression Testing Selection Techniques

The work is an attempt to develop an algorithm for reduction of test cases from a large test suite using genetic algorithm and bee colony optimizations that will reduce both time and effort and produce optimal results.

The paper is organized as follows. Section II pertains to related works and research in this specific area. Section III discusses how the genetic algorithm works. Section IV talks about the artificial bee colony optimization. The proposed approach has been discussed in Section V and illustrated with an example. The application of proposed approach is given in Section VI. Section VII summarizes the conclusion and the future work.
II. RELATED WORK

Several algorithms based on genetic algorithm [20, 21] and swarm intelligence [22, 23], i.e., ant colony optimizations and bee colony optimizations, have been proposed for test case selection and prioritization from a large test suite. Shameri[24] and Pargas et al [25] applied GA for automatic test data generation in his thesis. A Strategy for using GA to automate branch and fault-based Testing [26] and automatic structural testing using genetic algorithms [27] is done by Jones et al. Lin and Yeh worked on GA for automatic test data generation based on path based testing [28]. An evolutionary approach is developed to dynamic test data generation by Anastasis and Andreas [29]. Harman et al proposed an approach to reduce the input domain using search based technique [30]. In fact, the genetic algorithm is also used to generate test data automatically [31]. A lot of work is done by researchers on optimization of test cases. Mala et al has developed a hybrid genetic algorithm based approach for quality improvement and optimization of test cases [32] and Eric et al analyzed the effect of fault detection of test set when its size is minimized [33]. The concept of Artificial Bee Colony algorithm was introduced by Karaboga [34-36]. Chong et al [37] applied honey bees foraging behavior model to the job scheduling problem. McCaffrey et al [38] generates pair wise test sets using a simulated bee colony algorithm. Mala et al [39] presented a new, non pheromone based test suite optimization approach inspired by the behavior of biological bees. Dahiya et al [40] presented an ABC algorithm based approach for automatic generation of structural software tests.

III. GENETIC ALGORITHM

A genetic algorithm (GA) is an optimization technique that can be applied to various problems, including those that are NP-hard [41]. GA is an adaptive search procedure that was introduced by John Holland [20] and extensively studied by Goldberg [21], De Jong [42] and many other researchers. It uses a “survival of the fittest” technique, where the best solutions survive and are varied until we get a good product [43].

To apply GA, two main requirements are to be satisfied:

(a) An encoding is used to represent a solution of the solution space, and

(b) An objective function i.e. a fitness function which measures the goodness of a solution [44].

The proposed technique makes use of genetic operations. The GA process consists of the various steps as shown in fig 2.

Fig. 2: GA Architecture

Encoding is done for the solution to the problem. Using fitness-based function like roulette wheel selection and tournament selection, initial population is chosen. The second generation population of solutions is generated from first generation using genetic operators like crossover and mutation. New population will be chosen and further take part in generating the next generation. The process is repeated until a termination condition is reached (i.e., the result has been found or, fixed number of generations reached).

In the proposed technique, crossover operation is applied at the second step of GA life cycle. This is the method of merging the information units of two individuals that will produce two more new children (information units). Here cutting of the two strings at the user crossover point and swapping the two. The outcome of this process is the new population.

Take two strings and perform a 2-point crossover on them.

```
10111      00101
11111      00001
```

The new population or strings generated after applying crossover are: 10111000101 and 1111100101.

IV. ARTIFICIAL BEE COLONY OPTIMIZATION

Artificial Bee Colony (ABC) algorithm is a swarm based meta-heuristic algorithm introduced by Karaboga in 2005 [22, 23], and simulates the foraging behavior of honey bees. BCO is an optimization tool which provides a population based search procedure, where the artificial bees modify the food positions with time [45]. The main aim of the bees’ is to locate the food source positions with high nectar amount [46]. The colony of bees in ABC algorithm comprises of three
groups of bees: employed bees, onlookers and scouts [45]. Employed bees forage in search of their food source and return to hive and perform a dance on this area. The employed bee who find abandoned food source becomes a scout and find a new food source again. Onlookers decided their food source depending upon the dances of employed bees.

A nectar source is selected by each bee by following a nest mate whose food source has already discovered. The bees dance on the floor area(i.e.) hive, on discovery of nectar sources. That is how bees convinced their nestmates to follow them. To get nectar, bees follow the dancers to one of the nectar areas. On collecting the nectar, bees follow the dancers to one of the nectar areas. The procedure by which the bee decides to follow a specific dancer, is not well understood but it is considered that “the recruitment among bees is always a function of the quality of the food source” .It is also noted that not all bees start foraging simultaneously.

V. PROPOSED APPROACH

In this paper, we proposed a new approach to reduce the cost of regression testing by test case suite reduction. The proposed technique is based on concepts of BCO and GA. The technique selects the set of test case from the available test suite that will cover all the faults detected earlier in minimum execution time. Here bees are used as agents who explore the minimum set of test cases. Half of the bees will initially start foraging with randomly selected test cases. Now bees will add new test cases on her explored path if adding of a test case increases its fault detection capacity. After adding one more test case, the bees return to their hive, and exchange the information based on GA. The crossover operation is used to exchange the information. The new set of test case produced after crossover is used by new bees to forage. The process is repeated till any of the bees has discovered a set of test cases that covers all faults detection and starts to perform waggle dance.

The prerequisite for the proposed algorithm is a test suite 'T' of ‘n’ test cases. The result is subset 'S', which consists of m test cases(m<n),such that the test cases are selected on the basis of maximum fault coverage capacity in minimum execution time.

The assumptions taken for the proposed algorithm is as follows:

- Given the original test suite, T={t1,t2……tn}.
- Set of all faults, F={f1,f2,…….fk}.
- Each test case {t1,t2....tn} in the original test suite covers some or all the faults from ‘F’.
- Zi is the execution time for running each test case ti , where. 1≤ i ≤ n.
- Each test case will be represented in binary form. Each test case is of ‘k’ bits (k is the total number of faults). Each bit of the test case depends upon the capacity of detecting that fault. Starting from the leftmost bit, the bit is 1 if it detects Fault F_{i} else 0 and so on.
- Number of artificial bees ‘B’ to search through the test case space is n (number of test cases).
- It is not necessary that all bees will fly at the same time. If we have ‘n’ bees, then ‘n/2’ bees will start flying initially.
- Initially number of faults covered by bees is zero.
- While foraging, bees try to cover more faults as much as possible.
- Bees while foraging, will add the test case only if it increases its number of faults covered capacity by the set of test cases, otherwise the bee searches for the other test case.

The steps for running the proposed algorithm are:

Step 1: Initially the ‘n/2’ starts foraging. All the bees will not initially start foraging.

Step 2: Each bee that have started forage will choose the test case randomly.

Step 3: The bees will select test case on the basis that adding test case will increase the fault detection capacity. This can be checked by ‘OR’ operation of Boolean algebra. The number of 1’s present in the result is the number of faults covered.

Step 4: The foraged bees will return to their hive and genetically exchange information by crossover method.

Step 5: Crossover will be performed between the bees whose total execution time is minimum and the bees with the next minimum execution time. We assume that the bees that discover a set of test case in minimum time will produce a new set of test case that will be executed in minimum time and helpful in future to predict better and optimum result.

Step 6: If the resulted test cases’s total execution time is less than the maximum execution time available subset
of test cases, the new bees will forage using those subset of test cases as its initial path.

Step 7: If the results produced after crossover does not produce new test cases or test cases which are superfluous, will not be considered. If both test sets produced after crossover will not produce new set, no new bee will forage.

Step 8: Now again the bees will choose the test case. Repeat From step 3 to 6 till any of the bee has explored a set of test cases that can cover all the test faults.

Step 9: As soon as the minimum set of test case are produced, bee started to perform a waggle dance announcing her victory. So the other bees can follow this test case path.

Step 10: As the number of iterations increase the system will give the optimum result.

Example:

Consider a test suite with test cases in it, covering a total of 10 faults. Test case selection selects a subset of test cases which will cover all the faults in minimum execution time. In this section we demonstrate the working and efficiency of proposed approach using the given example.

The regression test suite ‘T’ as given in Table 1, contains eight test cases {T1, T2, T3, T4, T5, T6, T7, T8}. The prerequisite for this example assumes the knowledge of the faults detected by T as shown in Table 2. Test case T1 can find four faults [F1,F3,F6,F9] in seven minutes, T2 finds two faults [F2,F8] in three minutes, and T3 finds three faults [F2, F5,F7] in 5 minutes. Test cases T4 and T5 find four and three faults in 5 and three minutes [F4,F6,F8,F9] and [F1,F6,F10] respectively. Test cases T6 and T7 find three faults in six and three minutes [F4,F5,F8] and [F1,F7,F8]. Test case T8 find two faults [F3,F10] in two minutes.

Table 1: Sample data

<table>
<thead>
<tr>
<th>Test Case</th>
<th>No. of faults covered</th>
<th>Execution Time(ET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>T2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>T4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>T5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T6</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>T7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T8</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Test cases, no. of faults covered, their execution time

The algorithm works as follows:

The binary representation of test cases is shown in Table 3. The bit id is 1 (high) if it covers the respective fault starting from the leftmost position, otherwise 0.

Table 3: Binary Representation of Test cases

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Binary Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1010010010</td>
</tr>
<tr>
<td>T2</td>
<td>0100000100</td>
</tr>
<tr>
<td>T3</td>
<td>0100101000</td>
</tr>
<tr>
<td>T4</td>
<td>0001010110</td>
</tr>
<tr>
<td>T5</td>
<td>1000100011</td>
</tr>
<tr>
<td>T6</td>
<td>0001100100</td>
</tr>
<tr>
<td>T7</td>
<td>1000001100</td>
</tr>
<tr>
<td>T8</td>
<td>0010000001</td>
</tr>
</tbody>
</table>

Table 4: Binary Representation of Faults

Let number of artificial bees in this example are B :{ B1, B2, B3, B4, B5, B6, B7, B8} that is equal to the number of test cases. Initially ‘n/2’ i.e.4 bees (8/2=4) B1, B2, B3, B4 will forage. The number of fault covered by each bee is zero initially. Suppose the test case chosen by B1, B2, B3, and B4 is T1, T2, T3, T4 respectively as shown in Table 4.
In Round II, bees will select those test cases that will increase its faults covering capacity. Suppose B1 selects T5. Check the number of faults by OR function of Boolean algebra.

**Table 4: Test cases selected by bees initially**

<table>
<thead>
<tr>
<th>BEES</th>
<th>Test Case Selected</th>
<th>Total Execution Time</th>
<th>No. of Faults Covered</th>
<th>Binary Representation of Fault Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>T1, T5</td>
<td>10</td>
<td>5</td>
<td>1010010011</td>
</tr>
<tr>
<td>B2</td>
<td>T2, T6</td>
<td>9</td>
<td>4</td>
<td>0101100100</td>
</tr>
<tr>
<td>B3</td>
<td>T3, T7</td>
<td>8</td>
<td>5</td>
<td>1100101100</td>
</tr>
<tr>
<td>B4</td>
<td>T4, T8</td>
<td>7</td>
<td>6</td>
<td>0011010111</td>
</tr>
</tbody>
</table>

So the number of faults detected by B1 by choosing \{T1, T5\} is 5. We will calculate this for all bees and the calculated values are shown in table 5.

**Table 5: Selecting TC by bees in Round II**

<table>
<thead>
<tr>
<th>BEES</th>
<th>Test Case Selected</th>
<th>Total Execution Time</th>
<th>No. of Faults Covered</th>
<th>Binary Representation of Fault Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>T1, T5, T2, T6</td>
<td>10</td>
<td>7</td>
<td>1110010111</td>
</tr>
<tr>
<td>B2</td>
<td>T2, T6, T5</td>
<td>13</td>
<td>7</td>
<td>1101110101</td>
</tr>
<tr>
<td>B3</td>
<td>T3, T7, T8</td>
<td>10</td>
<td>7</td>
<td>1110110110</td>
</tr>
<tr>
<td>B4</td>
<td>T4, T8, T7</td>
<td>10</td>
<td>8</td>
<td>1011011111</td>
</tr>
<tr>
<td>B5</td>
<td>T3, T8, T1</td>
<td>14</td>
<td>8</td>
<td>1110111110</td>
</tr>
<tr>
<td>B6</td>
<td>T4, T7, T3</td>
<td>13</td>
<td>8</td>
<td>1011111110</td>
</tr>
</tbody>
</table>

**Table 6: Foraged BEES and their execution time**

<table>
<thead>
<tr>
<th>BEES</th>
<th>Test Case Selected</th>
<th>Total Execution Time</th>
<th>No. of Faults Covered</th>
<th>Binary Form of Fault Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>T1, T5, T2, T6</td>
<td>10</td>
<td>7</td>
<td>1110010111</td>
</tr>
<tr>
<td>B2</td>
<td>T2, T6, T5</td>
<td>13</td>
<td>7</td>
<td>1101110101</td>
</tr>
<tr>
<td>B3</td>
<td>T3, T7, T8</td>
<td>10</td>
<td>7</td>
<td>1110110110</td>
</tr>
<tr>
<td>B4</td>
<td>T4, T8, T7</td>
<td>10</td>
<td>8</td>
<td>1011011111</td>
</tr>
<tr>
<td>B5</td>
<td>T3, T8, T1</td>
<td>14</td>
<td>8</td>
<td>1110111110</td>
</tr>
<tr>
<td>B6</td>
<td>T4, T7, T3</td>
<td>13</td>
<td>8</td>
<td>1011111110</td>
</tr>
</tbody>
</table>

In Round III, the test case chosen by bees (Table 8) are as follows:

**Table 7: Selecting TC by bees in Round III**

Now all the six bees will return to their place after selecting test cases and genetically do a crossover operation for exchanging information. In this round B3 (ET: 10) and B4 (ET: 10) do a crossover operation.

<table>
<thead>
<tr>
<th>BEES</th>
<th>Test Case Selected</th>
<th>Total Execution Time</th>
<th>No. of Faults Covered</th>
<th>Binary Representation of Fault Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>T1, T5, T2</td>
<td>13</td>
<td>7</td>
<td>1110010111</td>
</tr>
<tr>
<td>B2</td>
<td>T2, T6, T5</td>
<td>12</td>
<td>7</td>
<td>1101110101</td>
</tr>
<tr>
<td>B3</td>
<td>T3, T7, T8</td>
<td>10</td>
<td>7</td>
<td>1110110110</td>
</tr>
<tr>
<td>B4</td>
<td>T4, T8, T7</td>
<td>10</td>
<td>8</td>
<td>1011011111</td>
</tr>
<tr>
<td>B5</td>
<td>T3, T8, T1</td>
<td>14</td>
<td>8</td>
<td>1110111110</td>
</tr>
<tr>
<td>B6</td>
<td>T4, T7, T3</td>
<td>13</td>
<td>8</td>
<td>1011111110</td>
</tr>
</tbody>
</table>

After applying crossover the new test cases produced are \{T3, T8, T7\} and \{T4, T7, T8\} whose total execution time is 10 and 10 and faults covered are 7 and 8. Since the results produced generate no new set of test cases so no new will forage.

Now, Round IV starts. In this Round, the test case chosen by bees in Table 9 are as follows:

**Table 8: Selecting TC by bees in Round III**

<table>
<thead>
<tr>
<th>BEES</th>
<th>Test Case Selected</th>
<th>Total Execution Time</th>
<th>No. of Faults Covered</th>
<th>Binary Form of Fault Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>T1, T5, T2, T6</td>
<td>19</td>
<td>9</td>
<td>1111110111</td>
</tr>
<tr>
<td>B2</td>
<td>T2, T6, T5</td>
<td>17</td>
<td>8</td>
<td>1101111101</td>
</tr>
<tr>
<td>B3</td>
<td>T3, T7, T8</td>
<td>15</td>
<td>10</td>
<td>1111111111</td>
</tr>
</tbody>
</table>
As we have seen that B3 and B6 has covered all the faults in 15 units. So test case set of \{T3, T4, T7, T8\} is the first minimum set of test case produced. To produce more optimal efficient results the above process should be iterated.

VI. APPLICATION OF PROPOSED APPROACH

Software practitioners may use the technique developed to reduce time and effort required for selection of test cases from a large test suite. This approach may start to produce better results. But as the iterations proceeds, the system will produce optimum results. This approach proved very beneficial during regression testing. The test cases selected, will cover all the faults. Hence the proposed approach may prove to be useful in real life situation.

VII. CONCLUSION AND FUTURE WORK

We have proposed test case selection approach from a large test suite using hybrid technique based on genetic algorithms and bee colony optimizations. This approach has been tested for several examples. One of these examples has been shown in this paper. The technique developed using this approach identifies and reduces the test data. The approach provides better results in the initial iteration of the whole process. It provides positive feedback and hence it can lead to better solutions in optimum time.

Issues of future research include automation of the technique and applying it on large and complex software. We also aim to compare it to ant colony optimizations algorithms and genetic algorithms.

REFERENCES


