**A Technique for Generating Test Data using**

**Genetic Algorithms**

1st Author

1st author's affiliation  
1st line of address  
2nd line of address  
Telephone number, incl. country code

1st author's E-mail address

2nd Author

2nd author's affiliation  
1st line of address  
2nd line of address  
Telephone number, incl. country code

2nd E-mail

3rd Author

3rd author's affiliation  
1st line of address  
2nd line of address  
Telephone number, incl. country code

3rd E-mail

**ABSTRACT**

Automatic path coverage test data generation is an undecidable problem and genetic algorithm (GA) has been used to test data generation since 2000. This paper presents a method for optimizing GA efficiency by identifying the most critical path clusters in a test function. We do this by using the static program analysis to find all the paths having the path conditions with low probability in generating coverage data, then basing on these path conditions to adjust the procedure of generating new populations in GA. The proposed approach is also applied some test functions. Experimental results show that improved GA which can generate suitable test data have higher path coverage than the traditional GA.

**Categories and Subject Descriptors**

D.3.3 [**Programming Languages**]: Language Contructs and Features – *abstract data types, polymorphism, control structures.* This is just an example, please use the correct category and subject descriptors for your submission*.* The ACM Computing Classification Scheme: <http://www.acm.org/class/1998/>

**General Terms**

Your general terms must be any of the following 16 designated terms: Algorithms, Management, Measurement, Documentation, Performance, Design, Economics, Reliability, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Legal Aspects, Verification.

**Keywords**

Generic algorithm; path coverage testing; automatic test data generation.

# INTRODUCTION

Software quality becomes more important than ever and software testing is the most significant measure for it. However, software testing is very laborious and costly due to the fact that it is mostly made by manual [1]. In general, software testing accounts for approximately 50 percent of the elapsed time and more than 50 percent of the total cost in software development [2]. Thus, automated software testing is a promising way to cut down time and cost.

Automatic structural test data generation is a crucial problem in software testing automation and its implementation cannot only significantly improve the effectiveness and efficiency but also reduce the high cost of software testing. We focus on path coverage test data generation in respect that various structural test data generation problem can be transformed into a path coverage test data generation problem. Furthermore, path coverage testing strategy can detect almost 65 percent of errors in program under test [3].

Although path coverage test data generation is an undecidable problem [4], researchers still attempt to develop various methods and have made some progress. These methods can be classified into two types: static methods and dynamic methods.

Static methods include symbolic execution [5] and domain reduction [6, 7] etc. These methods suffer from a number of problems when it handles indefinite loops, array, procedure calls and pointer references [8].

Dynamic methods include random testing, local search approach [9], goal-oriented approach [10], chaining approach [11] and evolutionary approach [8, 12-14]. Since values of input variables are determined when programs execute, dynamic test data generation can avoid those problems with that static methods are confronted.

As a robust search method in complex spaces, genetic algorithm (GA) was applied to test data generation in 1992 [12] and evolutionary approach has been a burgeoning interest since then. Related works [8, 15, 16] indicate that GA-based test data generation outperforms other dynamic approaches e.g. random testing and local search.

As far as we know, even though GA-based test data generation already proved its efficiency in generating test data for dynamic approaches, it still has to face difficulties when applying to path coverage, which are for the test function having test path with low probability in generating coverable test data, traditional GA cannot generate test data which can cover these difficult paths.

Consider below program under test:

1 void Example(int x, int y, double z){

2 boolean flag = y > 1000;

3 // ...

4 if (x + y == 1024)

5 if (flag)

6 if (Math.cos(z)- 0.95 < Math.exp(z))

7 // target branch

8 // ...

9 }

From experience the traditional GA cannot generate test data which can cover the first condition in line 4.

This paper gives the proposal to improve traditional GA in generating test data which can cover all the paths in a test function. It combines static program analysis with GA. The static program analysis step is applied to find out paths of the test function which are difficult to be covered. In this paper, the difficult path mean the path contains equal conditions which is difficult generate test data to cover. After that, basing on the path condition of these difficult paths, adjusts the procedure of generating new populations in GA.

This paper is organized as follows: Section 2 gives some theoretical background to understanding this research. Section 3 summarizes some related works, and Section 4 presents the proposed approach in detail. Section 5 shows the experimental results and discussion. Section 6 concludes the paper

# BACKGROUND

This section describes to the two content is the theoretical background for the proposal of this paper, path coverage test data generation as an optimization problem and genetic algorithm.

## Path coverage test data generation as an optimization problem

To make use of genetic algorithm, a path coverage test data generation problem requires being transformed into an optimization problem.

Firstly, program under test should be represented by its control flow graph (CFG). A CFG is a directed graph which can be denoted as G = (N, A, s, e) where N is a set of nodes, A is a set of edges; s and e are unique entry and unique exit node respectively. Each decision node is associated with a branch predicate, which is a logical expression. The edges leaving decision nodes are labeled with true or false values for corresponding branch predicate. To cause a path to be covered during execution, it is necessary to find appropriate values for the input variables that satisfy related branch predicates. A simple way is Korel’s branch distance function [9] based approach. For example, if a branch predicate C is (a == b), then the branch distance function f(C) = abs(a - b). So, to achieve a desired branch is transformed to search input vector that minimize its branch distance function. Table 1 gives some common used branch distance functions. To achieve a desired path P, we can define F(P) as the sum of all related branch distance functions. Consequently, generating path coverage test data can be transformed into searching input vector that can minimize F(P).

Table 1. Korel’s branch distance function

|  |  |  |
| --- | --- | --- |
| **No** | **Branch** | **Branch distance function** |
| 1 | a = b | f(C) = abs(a - b) |
| 2 | a ≠ b | f(C) = k |
| 3 | a < b | f(C) = (a - b) + k |
| 4 | a ≤ b | f(C) = (a - b) |
| 5 | a > b | f(C) = (b - a) + k |
| 6 | a ≥ b | f(C) = (b - a) |
| 7 | C1 ∧ C2 | f(C) = min(f(C1), f(C2)) |
| 8 | C1 ∨ C2 | f(C) = f(C1) + f(C2) |

In Korel function, k is the smallest step between 2 operands in the condition. In this paper, because all operand types are double, so to simplify we assumse k = 0.

## Generic algorithm

The basic concepts of genetic algorithm (GA) were developed by Holland [17]. GA is commonly applied to a variety of problems involving search and optimization. GA search methods are rooted in the mechanisms of evolution and natural genetics. GA draw inspiration from the natural search and selection processes leading to the survival of the fittest individuals. GA generates a sequence of populations by using a selection mechanism, and use crossover and mutation as search mechanisms.

The principle behind GA is that they create and maintain a population of individuals represented by chromosomes (essentially a character string analogous to the chromosomes appearing in DNA). These chromosomes are typically encoded solutions to a problem. The chromosomes then undergo a process of evolution according to rules of selection, crossover and mutation.

Each individual in the environment (represented by a chromosome) receives a measure of its fitness in the environment. Reproduction selects individuals with high fitness values in the population, and through crossover and mutation of such individuals, a new population is derived in which individuals may be even better fitted to their environment. The process of crossover involves two chromosomes swapping chunks of data (genetic information) and is analogous to the process of sexual reproduction. Mutation introduces slight changes into a small proportion of the population and is representative of an evolutionary step. The structure of a traditional GA is given below.

1 Traditional Genetic Algorithm (){

2 initialize population;

3 evaluate population;

4 while (stopping criteria not reached){

5 select solutions for next population;

6 perform crossover and mutation;

7 evaluate population;

8 }

9 }

The algorithm will iterate until the population has evolved to form a solution to the problem, or until a maximum number of iterations have taken place (suggesting that a solution is not going to be found given the resources available).

# RELATED WORK

The path coverage literature using GA started with Lin and Yeh [18] in 2000. They extended Jones et al.'s work [19] from branch coverage to path coverage. The ordinary (weighted) Hamming distance was extended to handle different ordering of target paths that have the same branches. The fitness function is called SIMILARITY, which computes similar items with respect to their ordering within two different paths between actual executed path and the target path. Only one program was used to test the approach, i.e. simple triangle classifier. They reported that the approach outperformed random search. However, in this method, evaluation function must be called many many times in order to generate the test data for the most difficult path to be covered. In addition, because their work only used dynamic analysis so above program under test Example cannot be covered all test paths.

Bueno et al. [20] proposed an approach that utilizes control and data flow dynamic information to achieve path coverage testing using GA. In addition, the work also tackled the detection of infeasible paths by monitoring the progress of evolutionary search. The fitness function was formulated by number of coincidence branches and the normalized branch predicate value at which the actual executed path starts to deviate from the target path. Six small test programs were used to validate the approach, with 10 repetitions each to minimize random variations. Two execution modes were used, i.e. one with initialized population and the other with a random initial population. The experiment results were promising.

In 2003, Hermadi and Ahmed [21] presented evolutionary test data generation for path testing using multiple paths. Prior to this work, almost all of the evolutionary test data generators only sought to cover a single target path at a time. The fitness function used the number of matching branches and branch predicate values using Korel's fitness function [9]. It also considered path traversal techniques, neighborhood influence, weighting, and normalization. Three small programs were used to validate the approach: minimum-maximum finder, triangle classifier, and a combination of both of them. Results were more effective and efficient by tackling multiple paths at a time.

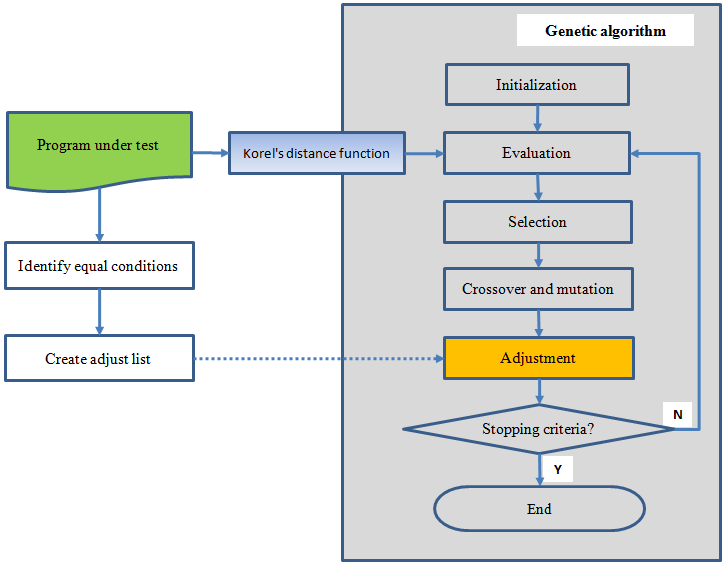
In 2008, Ahmed and Hermadi [23] extended their work of 2003 [21]. The extensions were adding a rewarding scheme and using a more efficient test data generator. A total of 32 fitness function variations were tested empirically and analyzed to determine which the best was. There were 7 test programs used in the experiments. The results demonstrated that the approach was better compared to other existing work.

In the same year, Chen and Zhong [24] developed a multi-population genetic algorithm for path testing. This work has been improving GA-based path testing as described in Section 2.2. The work reported that the proposed approach outperformed a traditional genetic algorithm based approach, using the triangle classifier as the test function. Similar to our approach, this paper also targets finding the test data to cover path conditions of the most difficult path to be covered in test function. As it approached the parallel processing, test data generating time is better than traditional GA, however the number of test data generation is still high (requires 21,073 test data by average).

In [25], Srivastava P.R and Kim T have presented a method for optimizing software testing efficiency by identifying the most critical path clusters in a program. The software under test is converted into a CFG. Weights are assigned to the edges of the CFG by applying 80-20 rule. 80 percentage of weight of incoming credit is given to loops and branches and the remaining 20 percentage of incoming credit is given to the edges in sequential path. The summation of weights along the edges comprising a path determines criticality of path. Higher the summation more critical is path and therefore must be tested before other paths. In this way by identifying most critical paths that must be tested first, testing efficiency is increased.

# PROPOSED APPROACH

This section describes details of our proposed approach, to test data generation using GA. In order to generate test data which can cover the paths having the lowest coverable probability, we propose 2 step approaches as in the above flow chart:



**Fig. 1. Flow chart of our proposed approach**

## Perform static program analysis

Mục đích của step này tạo ra một danh sách các điều kiện so sánh bằng từ một program under test. This list is used as conditions of adjustment for GA in the next step. Để tạo được danh sách này chúng tôi đã thực hiện các bước sau đây:

### Create a class to contain an equal condition

Using a list of below class to contain all equal conditions of a program under test.

class Adjust

{

public int index1; // index of input parami

public boolean bIsValue; // compare between parami

// and paramj or not

public int index2; // index of input paramj

public double comparevalue;

public String op; // operator between input params

}

### Parse equal condition

Giả sử program under test có *n* input parameters *param*i, *i* = 1~*n*. Các loại equal condition của program under test mà chúng tôi có thể tự động detect để cho vào danh sách là:

* *param*i == *param*j, *i*, *j* = 1~*n*
* *param*i op *param*j == *value*, *i, j* = 1*~n*, op = {+, -, ×, ÷}
* *param*i == *param*j op *param*j == *param*k, *i, j, k = 1~n*, op = {&&, ||}

Quay lại với program under test example đã đề cập ở section 1, điều kiện so sánh bằng ở line 4 sẽ được phân tích và lưu trữ ở cấu trúc Adjust như sau:

Adjust.index1 = 1; // input param x

Adjust.bIsValue = false; // x is compared to y, not

// compared to a value

Adjust.index2 = 2; // input param y

Adjust.comparevalue = 1024;

Adjust.op = “+”;

## Execute GA

To automatically generate test cases, using GA with below procedures:

### Representation

Depend on the type of input parameters of test function, GA uses a double or integer vector as a chromosome *chrom* = (*x*1, *x*2… *xn*) to represent values of the input variables. The length of the vector depends on the required precision and the domain length for each input variable.

### Initial population

At first, it needs to identify a fixed popsize number is the number of chromosome in a population (called *popsize*) also maximum population generation for each time to run GA (called *maxgen*). Then initialize random values for all chromosomes in the first population.

### Evaluation function

Để áp dụng được Korel’s branch distance function thì cần phải insert instrumented code vào program under test và sử dụng chính program under test này như là evaluation function of GA. For example, with above program under test Example, instrumented code sẽ được insert vào original code before line 4, 5, 6 as below:

1 double Example(int x, int y, double z){

2 boolean flag = y > 1000;

3 // ...

bouble ret = Math.abs((x + y) - 1024);

4 if (x + y == 1024) {

ret += (1000 - y);

5 if (flag){

ret += (Math.cos(z) - 0.95) - Math.exp(z);

6 if (Math.cos(z)- 0.95 < Math.exp(z))

7 // target branch

}

}

8 // ...

return ret;

9 }

### Selection

A selection scheme is applied to determine how individuals are chosen for mating based on their fitness. Fitness can be defined as a capability of an individual to survive and reproduce in an environment. Selection generates the new population from the old one, thus starting a new generation. Each chromosome is evaluated in present generation to determine its fitness value. This fitness value is used to select the better chromosomes from the population for the next generation.

### Crossover and mutation

After selection, the crossover operation is applied to the selected chromosomes. It involves swapping of values of vector *x* = (*x*1, *x*2,…, *xn*) between two chromosomes. This process is repeated with different parent chromosomes until the next generation has enough chromosomes. After crossover, the mutation operator is applied to a randomly selected subset of the population. Mutation alters chromosomes in small ways to introduce new good traits. It is applied to bring diversity in the population.

### Adjustment

Mục đích của thủ tục điều chỉnh là giúp cho GA nhanh chóng sinh ra được test data mà có thể cover được toàn bộ test path của program under test. So that after executing the mutation, based on list of equal conditions which are contained in list adjust, we need to adjust the values of each chromosome in the population. The adjustment will be executed as follows:

for each adjust[i] in the adjust list

if (adjust[i].bIsValue)

chrom.x[adjust[i]] = adjust[i].comparevalue

else

switch case of (adjust[i].op (+/-/\*/÷))

chrom.x[adjust[i].index1]=

adjust[i].comparevalue (+/-/\*/÷)

chrom.x[adjust[i].index2]

# EXPERIMENTAL RESULTS

Ở phần này chúng tôi sẽ trình bày về kết quả thực hiện của improved GA so với traditional GA.

## Programs under test

Ngoài program under test example đã trình bày ở trên, để minh họa cho kết quả cải tiến GA chúng tôi thực hiện thêm trên 2 program test under QuadEqua2 và Tritype như sau:

### QuadraticEquation2 function

This function finds all roots of a quadratic equation with 3 coefficients a, b and c as input parameters. It has 2 equal conditions which are stored in adjustment list: [a == 0] and [b == 0].

void QuadEqua2(double a, double b, double c) {

if (a == 0) {

if (b != 0) {

root1 = (-c)/b;

}

else {

// no any root

}

return;

}

if (((b\*b)-(4\*a\*c)) < 0) {

// no any root

return;

}

else {

if (((b\*b) - (4\*a\*c)) == 0) {

root1 = (-b)/(a\*2);

}

else {

root1 = (-b + Math.sqrt(((b\*b) - (4\*a\*c))))/(2\*a);

root2 = (-b - Math.sqrt(((b\*b) - (4\*a\*c))))/(2\*a);

}

}

}

### Tritype function

This function [22] determines whether three given numbers that represent three lengths on a plane form a scalene, isosceles, equilateral, or not a triangle. It has 2 equal conditions which are stored in adjustment list: [a == b] and [b == c].

void Tritype(double a, double b, double c) {

if (a + b > c && a + c > b && b + c > a) {

if (a == b && b == c) {

type = 3; // Equilateral

}

else {

if (a == b || b == c || c == a) {

type = 2; // Isosceles

}

else {

type = 1; // Scalene

}

}

}

else {

type = -1; // Not a triangle

}

}

## GA parameters setting

Parameter settings of traditional GA and improved GA are as following:

* Length of the chromosome: 3
* Selection method: random
* Two-point crossover probability (pc): 0.5
* Mutation probability (pm): 0.1
* Stopping criteria: all test target paths are covered

Each program under test still requires other parameters below:

Table 2. GA parameter setting for each program

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Program** | **Type** | **Range** | **Max gen** | **Pop size** |
| Example | double | [-10000,10000] | 150 | 250 |
| QuadEqua2 | double | [-10000,10000] | 150 | 250 |
| Tritype | double | [0,10000] | 150 | 250 |

* Type: type of input variables
* Range: range of input variables
* Maxgen: maximum population generation for each time to run GA
* Popsize: number of chromosome for each population

## Results

Kết quả sinh test data của improved GA so với traditional GA được thể hiện ở trong bảng sau. Chúng ta sẽ đánh giá theo 2 tiêu chí là số lượng các test path được phủ và số lần phải thực hiện sinh test data.

### Test path coverage

Tiêu chí này được đánh giá dựa trên việc GA có khả năng sinh ra được test data mà có thể coverage được bao nhiêu test path của program under test.

Table 3. Test path coverage

|  |  |  |  |
| --- | --- | --- | --- |
| **Program** | **Feasible path** | **Traditional GA** | **Improved GA** |
| Example | 4 | 2 | 4 |
| QuadEqua2 | 5 | 2 | 4 |
| Tritype | 4 | 3 | 3 |

From Table 4, we can see that for 2 program under test (Example and QuadEqua2) thì improved GA có thể sinh ra được số test path coverage cao hơn so với traditional GA.

### Test data generation counts

Tiêu chí này đánh giá dựa trên số lần sinh test data để có thể phủ được toàn bộ feasible paths trong program under test. Trong trường hợp không thể sinh được test data để phủ được toàn bộ feasible paths thì test data generation counts = *maxgen* x *popsize*.

Table 4. Test data generation counts

|  |  |  |  |
| --- | --- | --- | --- |
| **Program** | **Feasible path** | **Traditional GA** | **Improved GA** |
| Example | 4 | 37500 | 346 |
| QuadEqua2 | 5 | 37500 | 37500 |
| Tritype | 4 | 25847 | 508 |

From Table 4, we can see that improved GA chỉ cần sử dụng số lần sinh test data ít hơn để sinh ra được test data phủ được toàn bộ các test path của program under test.

In addition, comparing to [18, 24], for program under test Tritype, the number of test data generation which can cover all 4 test paths is lower (508 times, comparing to 10000 times [18] or 21073 times [24]), proving that our improved GA is more effective.

# CONCLUSION

In software development life cycle, software testing is one of the critical phases. So generation of test data automatically is a key step which has a great influence on code coverage in software testing. In this paper, we have improved the GA in order to generate test data automatically for feasible target paths.

Our approaching method is to combine the static analysis in order to find path conditions of difficult path to be covered in test functions, and then adjust the procedure of generating the new population in GA in order to generate test cases which can cover these paths.

The experimental results on these test functions shows that improved GA can generate test data can cover path having path conditions which cannot be covered by test data generated from normal GA.

Besides, comparing to the current methods [18, 24], our proposal is more effective as it can more quickly generate test data which can cover paths that other methods cannot.

# REFERENCES

1. B. Antonia, "Software Testing Research: Achievements, Challenges, Dreams," in 2007 Future of Software Engineering: IEEE Computer Society, 2007.
2. G. J. Myers, The Art of Software Testing, 2nd edition: John Wiley & Sons Inc, 2004.
3. B. W. Kernighan and P. J. Plauger, The Elements of Programming Style, McGraw-Hill, Inc, New York, NY, USA, 1982.
4. E. J. Weyuker, The applicability of program schema results to programs, International Journal of Parallel Programming, vol. 8, 387-403, 1979.
5. C. K. James, A new approach to program testing, in Proceedings of the international conference on Reliable software Los Angeles, California: ACM, 1975.
6. T. Y. Chen, T. H. Tse, and Z. Zhiquan, Semiproving: an integrated method based on global symbolic evaluation and metamorphic testing, in Proceedings of the 2002 ACM SIGSOFT international symposium on Software testing and analysis Roma, Italy: ACM, 2002.
7. S. Nguyen Tran and D. Yves, Consistency techniques for interprocedural test data generation, ACM SIGSOFT Software Engineering Notes, vol. 28, 108-117, 2003.
8. G. M. C C Michael, M Schatz, Generating software test data by evolution, IEEE Transactions on Software Engineering, vol. 27, 1085-1110, 2001.
9. B. Korel, Automated software test data generation, IEEE Transactions on Software Engineering, vol. 16, 870-879, 1990.
10. B. Korel, Dynamic method for software test data generation, Software Testing, Verification & Reliability, vol. 2, 203-213, 1992.
11. B. Korel, Automated test data generation for programs with procedures, in Proceedings of the 1996 ACM SIGSOFT international symposium on Software testing and analysis San Diego, California, United States: ACM, 1996.
12. S. Xanthakis, C. Ellis, C. Skourlas, A. Le Gall, S. Katsikas, and K. Karapoulios, Application of genetic algorithms to software testing (Application des algorithmes genetiques au test des logiciels), in Proceedings of 5th International Conference on Software Engineering and its Applications Toulouse, France, 625-636, 1992.
13. J. Wegener, A. Baresel, and H. Sthamer, Evolutionary test environment for automatic structural testing, Information and Software Technology, vol. 43, 841-854, 2001.
14. J. Wegener, B. Kerstin, and P. Hartmut, Automatic Test Data Generation For Structural Testing Of Embedded Software Systems By Evolutionary Testing, in Proceedings of the Genetic and Evolutionary Computation Conference: Morgan Kaufmann Publishers Inc., 2002.
15. S. Levin and A. Yehudai, "Evolutionary Testing: A Case Study, in Hardware and Software, Verification and Testing, 155-165, 2007.
16. W. Joachim, Andr, Baresel, and S. Harmen, Suitability of Evolutionary Algorithms for Evolutionary Testing, in Proceedings of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment: IEEE Computer Society, 2002.
17. J. H. Holland, Adaptation in Nature and Artificial Systems, Addison-Wesley, Reading, MA, 1975.
18. Jin-Cherng Lin and Pu-Lin Yeh, Using genetic algorithms for test case generation in path testing, In Proceedings of the 9th Asian Test Symposium 2000 (ATS '00), 241-246, December 2000.
19. Bryan F. Jones, Harmen-Hinrich Sthamer, and D.E. Eyres. Automatic structural testing using genetic algorithms, Software Engineering, 11(5):299-306, September 1996.
20. Paulo Marcos Siqueira Bueno and Mario Jino, Automatic test datageneration for program paths using genetic algorithms, International Journal of Software Engineering & Knowledge Engineering (IJSEKE), 12(6):691-709, 2002.
21. Irman Hermadi and Moataz A. Ahmed, Genetic Algorithm based test data generator, In Proceedings of the 2003 Congress on Evolutionary Computation (CEC), volume 1, pages 85-91, December 2003.
22. I. Hermadi, C. Lokan, R. Sarker, Dynamic stopping criteria for search-based test data generation for path testing, Information and Software Technology, 56 (4):395-407, April 2014.
23. Moataz A. Ahmed and Irman Hermadi, GA-based Multiple Paths Test Data Generator, Computers & Operations Research, 35:3107-3124, October 2008.
24. Yong Chen and Yong Zhong, Automatic path-oriented test data generation using a multi-population genetic algorithm, In Proceedings of the 4th International Conference on Natural Computation, 2008 (ICNC'08), volume 1, pages 566-570, October 2008.
25. Srivastava P. R and Kim T, Application of Genetic Algorithm in Software Testing, International Journal of Software Engineering and Its Applications, 3(4), 87-96, 2009.