大規模データセットと多種ミューテーション演算子を利用した欠陥限局に適するプログラム構造の再調査

久保 光生[1]　肥後 芳樹[1]　楠本 真二[1]

概要：プログラムに含まれる欠陥箇所を自動で推測する方法として，スペクトラムに基づく欠陥限局（SBFL）がある．SBFL では，各テストケースの成否と実行経路情報をもとに欠陥箇所を特定する．同じ機能を持つプログラムでも，その構造によって SBFL の欠陥限局精度には差が生じる．よって，プログラム構造を SBFL に適する形に変更することで，機能を保ちつつ SBFL の精度向上が期待される．先行研究では SBFL に適するプログラム構造を発見するために，SBFL スコアが提案された．SBFL スコアはプログラムが SBFL にどの程度適しているかを評価する指標の 1 つである．また，先行研究では，同じ機能を持つが構造の異なるプログラムペアを対象として SBFL スコアを計測し，SBFL に適したプログラム構造を得ている．しかし，実験対象のプログラム数が 10 個，ミューテーション演算子が 11 種類と少ないことが課題であった．そこで本研究では，実験対象のプログラム数を約 36 倍，ミューテーション演算子の種類数を約 2.5 倍にして実験を行った．実験の結果，新たに SBFL に適したプログラム構造が 4 つ特定できた．

1. はじめに

ソフトウェア開発において，デバッグは多くの労力とコストを必要とする作業である．ソフトウェア開発に必要なコストのうち，半分以上をデバッグ作業が占めているという報告もある [1]．そのため，デバッグを支援するための研究が盛んに行われている．デバッグ支援に関する研究分野の 1 つに欠陥限局がある．欠陥限局とは，プログラム中の欠陥箇所を推測する技術である．中でも近年，スペクトラムに基づく欠陥限局 (Spectrum-Based Fault Localization, 以降 SBFL) に関する研究が盛んに行われている [2]．SBFL では，テストスイートに含まれる各テストケースがどの文を実行し，どの文を実行しなかったかという情報を用いて，プログラムの欠陥と疑われる箇所を自動的に特定する．SBFL の基本的なアイデアは，多くの失敗テストで実行される文ほど欠陥である可能性が高く，多くの成功テストで実行される文ほど欠陥である可能性が低いと判断することである．

SBFL の精度は，欠陥自体の性質やテストの内容など，様々な要因に左右される [3]．その中でも，佐々木らによる先行研究 [4] は，プログラム構造に着目した．この先行研究では，プログラムがどの程度 SBFL に適するかを表す SBFL 適合性が提案されている．また，SBFL 適合性の評価指標として，SBFL スコアが提案されている．SBFL スコア計測の基本的なアイデアは，すべてのテストケースを通過するプログラムについて，ミューテーションテストを用いて，様々な箇所に意図的に欠陥を発生させることである．これらの欠陥が SBFL によってどの程度正確に特定できたか計測することにより，元のプログラムの SBFL 適合性を評価できる．先行研究では同じ機能を持つが構造の異なる 5 組のプログラムペアを用いて SBFL スコアを計測することにより，SBFL に適するプログラム構造が調査された．しかし，計測対象のプログラム数が 10 個と少なく，SBFL に適するプログラム構造が十分に明らかになったとはいえない．また，プログラムに意図的に欠陥を発生させるためのミューテーション演算子の種類数が 11 個と少なく，多くのプログラムで十分な数のミュータントを生成できない．この状況で測定された SBFL スコアは，SBFL 適合性の評価指標として適切ではない．そこで本研究では，実験対象として 361 個のプログラムを用いて SBFL スコアを計測することにより，SBFL に適する新たなプログラム構造の発見を目指す．また，ミュータントの数を増加させるために，16 種類の新たなミューテーション演算子を定義する．これにより，合計で 27 種類のミューテーション演算子をプログラムに適用する．

ミューテーション演算子の追加の結果，約 98.6% の実験対象プログラムで，ミュータントの数の増加が確認できた．また，実験対象プログラムの SBFL スコアを計測し，その結果を目視で確認することにより，SBFL 適合性が高

[1] 大阪大学大学院情報科学研究科 大阪府吹田市

いプログラム構造が新たに 4 つ特定できた．

2. 準備

2.1 スペクトラムに基づく欠陥限局（SBFL）

デバッグを支援する技術の 1 つに，欠陥限局がある．欠陥限局とは，プログラム中の欠陥箇所を推測する技術である．テストを用いた自動的な欠陥限局方法の 1 つに，スペクトラムに基づく欠陥限局 (Spectrum-Based Fault Localization, SBFL) がある．SBFL におけるスペクトラムとは，テストスイートの実行時にどの文が実行されたかという実行経路情報である．SBFL の基本的なアイデアは，多くの失敗テストで実行される文ほど欠陥である可能性が高く，多くの成功テストで実行される文ほど欠陥である可能性が低いと判断することである．

SBFL による欠陥箇所の特定方法について説明する．まず，すべてのテストケースを実行し，テストの成否と実行経路情報を記録する．次に，これらの情報を利用して，疑惑値と呼ばれる，欠陥である可能性の高さを示す値を文ごとに算出する．疑惑値の算出方法には様々あるが，Abreu らが SBFL で用いられる計算式の有効性を評価した結果，Ochiai の計算式 [5] が最も優れていると結論づけている [6]．

Ochiai の計算式における疑惑値 $susp(s)$ の算出方法を式 (1) に示す．ここで，$fail(s)$ は文 s を実行した失敗テストの数，$pass(s)$ は文 s を実行した成功テストの数，$totalFail$ は失敗テストの総数である．

$$susp(s) = \frac{fail(s)}{\sqrt{totalFail \times (fail(s) + pass(s))}} \qquad (1)$$

$susp(s)$ をすべての文 s に対して算出し，その値が高い文ほど，欠陥の原因箇所である可能性が高いと推測する．

なお，SBFL の疑惑値の算出においては，失敗テストの実行経路情報が最も重要な要素となる．失敗テストでどのような文が実行され，どのような文が実行されなかったかが，欠陥の原因箇所に対する大きな手がかりとなるためである．Ochiai の計算式では，分子が $fail(s)$ であることから，失敗テストの実行経路情報の重要さが見て取れる．

2.2 SBFL 適合性

先行研究 [4] で提案された SBFL 適合性について述べる．SBFL 適合性とは，プログラムが持つ品質特性の 1 つであり，プログラム自体が SBFL にどの程度適しているかを表す．プログラムの機能やテストスイートが同じでも，プログラム構造が異なれば SBFL を用いた欠陥限局の精度に違いが生じることがある．

プログラム構造の違いによる SBFL 適合性の変化について，例を用いて説明する．図 1 に示すプログラム (a) とプログラム (b) は，機能が同じだが，構造が異なる．図 1 に示すテストスイート (c) を用いて両方のプログラムに SBFL

プログラム (入力: a, b)	*susp*	t_1	t_2	t_3	t_4
s_1: `boolean result = false;`	0.50	✓	✓	✓	✓
s_2: `if (0 < a)`	0.50	✓	✓	✓	✓
s_3: `  result = true;`	0.00	✓	✓		
s_4: `if (0 <= b) //correct: 0 < b`	0.50	✓	✓	✓	✓
s_5: `  result = true;`	0.50	✓	✓	✓	✓
s_6: `return result;`	0.50	✓	✓	✓	✓
テスト結果：		P	P	P	F

(a) リファクタリング前

プログラム (入力: a, b)	*susp*	t_1	t_2	t_3	t_4
s'_2: `if (0 < a)`	0.50	✓	✓	✓	✓
s'_3: `  return true;`	0.00	✓	✓		
s'_4: `if (0 <= b) //correct: 0 < b`	0.71			✓	✓
s'_5: `  return true;`	0.71			✓	✓
s'_6: `return false;`	-				
テスト結果：		P	P	P	F

(b) リファクタリング後

テストケース	入力 (a, b)	期待値	実際の値
t_1:	(1, 1)	true	true
t_2:	(1, 0)	true	true
t_3:	(0, 1)	true	true
t_4:	(0, 0)	false	true

(c) テストスイート

図 1: プログラム構造の違いにより SBFL 適合性が変化する例

を実行すると，各文の疑惑値が算出される．プログラム (a) では，欠陥がある文と同じ疑惑値を持つ文が 4 個存在する．一方，プログラム (b) では，欠陥がある文と同じ疑惑値を持つ文は 1 個である．欠陥がある文と同じ疑惑値を持つ文が少ないほど，確認しなくてはならない文が少なくなるため，SBFL による欠陥限局の精度が高いといえる．よって，この場合，プログラム (b) の方が SBFL 適合性が高い．

2.3 SBFL スコア

SBFL 適合性と同時に先行研究 [4] で提案された SBFL スコアについて述べる．SBFL スコアは，SBFL 適合性の評価指標の 1 つとして提案された．SBFL スコアを計測するための基本的なアイデアは，ミューテーションテスト [7] の技術を活用し，与えられたプログラム中の可能な限り多くの箇所に意図的に変更を加え，人工的な欠陥を生成することである．これらの人工的な欠陥を SBFL でどの程度正確に特定できるか計測することにより，プログラム全体がどの程度 SBFL に適するかを測定できる．

2.3.1 SBFL スコアの計測方法

先行研究 [4] で提案された，SBFL スコアの計測方法の概要を示す．プログラム P の SBFL スコア計測時の入力は以下の 2 つである．

- ミュータント生成器 G
- テストスイート T

計測の大まかな流れを図 2 に示す．SBFL スコアの計測は以下の 3 つのステップで構成される．

(1) プログラムに対してミュータントを生成

(2) ミュータントに対して SBFL を実行

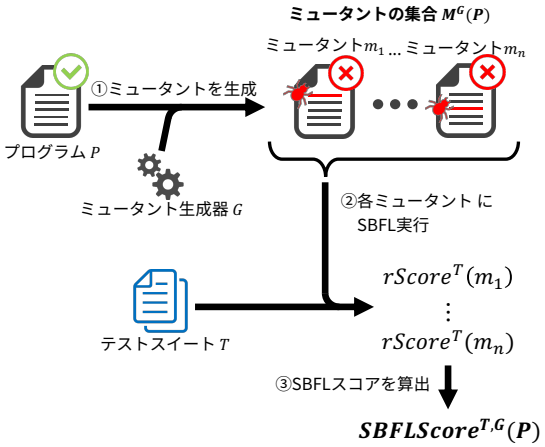


図 2: SBFL スコアの計測方法の概要

(3) SBFL スコアを算出

以降では，各ステップの詳細を述べる．

Step1. プログラムに対してミュータントを生成

対象プログラムとテストスイートを用意する．プログラムはテストスイート T に含まれるすべてのテストケースに通過することを前提とする．このプログラムに対して，ミュータント生成器を用いてミュータントを生成する．このとき，各ミュータントは元のプログラムに対して 1 箇所だけが変更されるように生成する．プログラムには変更可能な箇所が複数存在するため，ミュータントは複数生成される．ミュータント生成器 G によって生成されたすべてのミュータントの集合を $M^G(P)$ と定義する．

Step2. ミュータントに対して SBFL を実行

$M^G(P)$ に含まれる各ミュータントに対して，テストスイート T を用いて SBFL を実行し，文ごとの疑惑値を算出する．ここで，あるミュータント $m \in M^G(P)$ に含まれる各文 s について，以下を定義する．

- $susp^T(s)$：文 s の疑惑値
- $rank^T(s)$：文 s の疑惑値の順位
- $rScore^T(s)$：文 s の疑惑値の正規化順位

疑惑値の算出には Ochiai の計算式を用いる．疑惑値の順位は，疑惑値の高い順に文を並べた際に，欠陥が存在する文を発見するまでに最大で確認しなければならない文の総数とする．例えば，疑惑値 1.0 の文が 2 つ，0.8 の文が 1 つ存在する場合は，疑惑値 1.0 の文はどちらも 2 位と扱い，疑惑値 0.8 の文は 3 位とする．疑惑値の順位は，文の総数により異なる価値を持つ．例えば，10 個の文のうちの 10 位と，100 個の文のうちの 10 位とでは，後者の方が順位としての価値が高い．そこで，各文が文全体の中でどの程度上位に位置するかを表すため，順位を 0 以上 1 以下の範囲で線形に正規化する．文 s の正規化順位 $rScore^T(s)$ を以下の式 (2) の通り算出する．1 が最も価値が高く，0 が最も価値が低いことを表す．(2) において，$totalStatements^T$ はテストスイート T によって実行される文の数である．

$$rScore^T(s) = 1 - \frac{rank^T(s) - 1}{totalStatements^T - 1} \quad (2)$$

また，ミュータント m に含まれる欠陥の疑惑値の正規化順位を $rScore^T(m)$ と定義する．各ミュータントに含まれる欠陥は 1 箇所であるため，この値はミュータントごとに一意である．ミュータント m の欠陥を含む文を s^m_{fault} とすると，$rScore^T(m)$ は文 s^m_{fault} の疑惑値の正規化順位である．ミュータントの $rScore$ が高いほど，欠陥箇所を正確に特定できることを意味する．

$$rScore^T(m) = rScore^T(s^m_{fault})$$

Step3. SBFL スコアを算出

対象プログラムから生成されたすべてのミュータントの $rScore$ の平均をとることにより，SBFL スコアが算出される．なお，$|M^G(P)|$ は，生成されたミュータントの総数を表す．

$$SBFLScore^{T,G}(P) = \frac{1}{|M^G(P)|} \sum_{m \in M^G(P)} rScore^T(m)$$

2.4 SBFL スコアの信頼性

本研究では，SBFL スコアについて，信頼性という性質を新たを設ける．SBFL スコアの信頼性とは，SBFL スコアが元のプログラムの SBFL 適合性をどれほど正確に示しているかを表す性質である．SBFL 適合性は，プログラム中のあらゆる箇所の欠陥が SBFL でどの程度検出しやすいかを表すため，SBFL スコアを測定する際には，プログラム中のできるだけ多くの箇所に欠陥を生成することが必要である．生成されたミュータントの個数が多いほど，SBFL スコアが元のプログラムの SBFL 適合性を正確に表すため，SBFL スコアの信頼性は高まる．よって，SBFL スコアの信頼性は，プログラムから生成されたミュータントの個数から類推することができる．

2.5 先行研究の調査方法

先行研究 [4] では，Java で実装されたプログラムを対象として，プログラム構造の違いにより SBFL スコアがどのように変化するか確認するための実験が行われた．

2.5.1 実験対象プログラムとテストスイート

先行研究 [4] では，単一のメソッドで構成されるプログラムに対して SBFL スコアを計測する実験が行われた．実験対象は，同じ入力が与えられると同じ出力を返すが，プログラム構造が異なる 5 種類のメソッドペアである．なお，「プログラム構造」の正確な意味は 3.1 節で述べる．

5 種類のメソッドペアは，佐々木らによってリファクタリングを題材として作成された．リファクタリングには多

くのパターンが存在するが，Fowler によって「条件記述の単純化」に分類されたリファクタリングパターン [8] の中から選定された単一のメソッド内で完結する 5 種類のリファクタリングが題材となった．

また，各メソッドのテストスイートは佐々木らによって，以下を満たすように作成された．

- すべてのミュータントはいずれかのテストケースに失敗する．
- 条件網羅率が 100% になる．

2.5.2 ミュータント生成器

先行研究 [4] では，表 1 に示す 11 種類のミューテーション演算子が用いられた．これらは，オープンソースのミューテーションテストツールである PIT [9] の基本ミューテーション演算子を参考に実装された．PIT は，ミューテーションテストの分野でミュータントの生成に広く用いられている [10]．

2.6 先行研究の調査結果

先行研究 [4] での調査の結果，同一の条件分岐先で実行される文が少ないほど SBFL 適合性が高いことが明らかになった．また，SBFL 適合性を高めるプログラム構造の変換方法として，以下の 2 つが得られている．

- return 文を用いてメソッドを早期終了させる（図 3）
- 同一の条件分岐先の文の数が多い方から少ない方へ文を移動する（図 4）

3. 先行研究の課題とその解決策

先行研究 [4] には，以下の 2 つの課題が存在する．

- 実験対象プログラム数が少ない．
- ミューテーション演算子の種類数が少ない．

以下では，課題の詳細と，その解決策について述べる．

3.1 実験対象プログラム数が少ない

先行研究の実験対象プログラムは 5 組のメソッドペアに含まれる 10 個のメソッドであった．よって，調査結果の一般化可能性に欠ける．また，SBFL スコアが高くなるプログラム構造を十分に調査しきれていない．

表 1: 先行研究で用いられたミューテーション演算子

ミューテーション演算子	変換前	変換後
Conditional Boundary	`a<b`	`a<=b`
Increments	`n++`	`n--`
Invert Negatives	`-n`	`n`
Math	`a+b`	`a-b`
Negate Conditionals	`a==b`	`a!=b`
Void Method Calls	`method();`	`;`
Primitive Returns	`return 5;`	`return 0;`
Empty Returns	`return "str";`	`return "";`
False Returns	`return true;`	`return false;`
True Returns	`return false;`	`return true;`
Null Returns	`return object;`	`return null;`

```
1 int result = 0;
2 if (x > 0)
3   result = -10;
4 else if (y > 0)
5   result = -20;
6 else
7   result = -30;
8 return result;
```

(a) 変換前

```
1 if (x > 0)
2   return -10;
3 if (y > 0)
4   return -20;
5 return -30;
```

(b) 変換後

図 3: SBFL 適合性を高めるために return 文を用いてメソッドを早期終了する例

```
1 int result = 0;
2 int tmp = 0;
3 if (x > 0)
4   tmp = y * 2;
5 else
6   tmp = y * 3;
7 result = y + tmp;
8 return result;
```

(a) 変換前

```
1 int result = 0;
2 int tmp = 0;
3 if (x > 0)
4   tmp = y * 2;
5   result = y + tmp;
6 else
7   tmp = y * 3;
8   result = y + tmp;
9 return result;
```

(b) 変換後

図 4: SBFL 適合性を高めるために文を移動させる例

この問題の解決のために，本研究では，Java で実装された同機能メソッドを含むデータセット [11] を用いる．本データセットは，単一のメソッドからなるプログラムとテストスイートをそれぞれ 728 個ずつ含む．これらはオープンソースソフトウェアから収集された．また，メソッドは機能によって 276 組のグループに分類されている．各グループは，同機能メソッドを 2 個から 12 個含む．各メソッドのテストスイートは Evosuite [12] によって自動生成された．

なお，本研究で用いるデータセットの中には，記述方法は異なるが，同一のテストスイートを実行した際の実行経路が同じになるメソッドペアが含まれる可能性がある．「実行経路が同じ」とは，2 つのメソッドが以下の 3 つの条件を満たすことを指す．ここで，s_i は 1 つ目のメソッドの i 番目の文を，s'_i は 2 つ目のメソッドの i 番目の文を表す．

- メソッドに含まれる文の総数が同じである．
- テストスイートに含まれるすべてのテストケースにおいて，実行される文の総数 m が同じである．
- テストスイートに含まれるすべてのテストケースにおいて，$1 \leq i \leq m$ について，文 s_i における実行の有無が，文 s'_i における実行の有無と等しい．

実行経路が異なるメソッドペアの例を図 5 に，実行経路が同じであるメソッドペアの例を図 6 に示す．2 つのメ

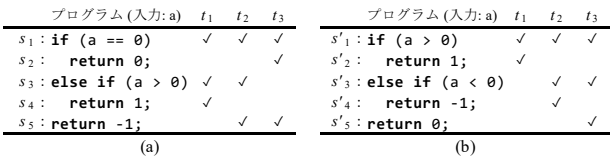


図 5: 実行経路が異なるメソッドペア

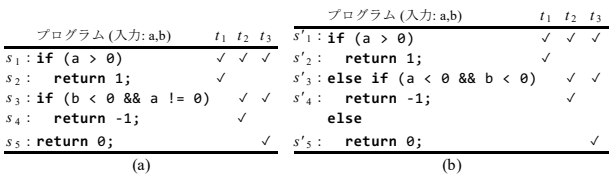


図 6: 実行経路が同じであるメソッドペア

ソッドの実行経路が同じになる場合，適用されるミューテーション演算子の違いのみによって，SBFL スコアに差が生じてしまう．このようなメソッドは実験対象として適切でないと考えたため，実験対象から除外する必要がある．よって，本研究では，プログラム構造を「同一のテストスイートを実行した際の実行経路」と定義する．記述方法が異なっても，同一のテストスイートを実行した際の実行経路が同じ場合には，プログラム構造は同じであるとみなす．

3.2 ミューテーション演算子の種類数が少ない

先行研究で使用されたミューテーション演算子は 11 種類であった．先行研究で実験対象となったプログラムは，これらのミューテーション演算子が多くの箇所に適用できるように実装されたため，十分な量のミュータントが生成できていた．しかし，それ以外のプログラムを用いる場合，十分な量のミュータントを生成できない場合がある．この状態で計測された SBFL スコアは，プログラムの SBFL 適合性を正確に表さないため，信頼性が低い．

この問題の解決のために，本研究では実験対象プログラムを目視し，既存のミューテーション演算子が適用不可能である箇所を抽出した．抽出した箇所に適用可能なミューテーション演算子として，新たに表 2 に示す 16 種類のミューテーション演算子を実装した[*1]．なお，新たに追加したミューテーション演算子は，既存のミューテーション演算子と同じミュータントを生成する可能性がある．例えば，プログラムに `return false;` という文がある場合，Change Boolean Literal と，True Return が同じミュータントを生成する．よって，同じミュータントは複数生成されないようにミュータント生成器を実装した．

また，SBFL スコアの信頼性の向上を検証するために，ミューテーション演算子の追加前と追加後のそれぞれについて，実験対象プログラムから生成されたミュータントの増加数を記録する．このとき，プログラムの規模を考慮する必要がある．なぜなら，プログラムの規模が大きくなると，ミュータントの個数は増加する傾向にあるからである．プログラムの規模に対するミュータントの増加数を評価するための指標として，IMPL (Increased Mutants Per LOC) を求める．IMPL は式 (3) で定義される．なお，論理 LOC とは，プログラムのうち，空行や括弧のみの行，コメントのみの行を除いた行数を指す．IMPL の値が大きいほど，SBFL スコアの信頼性が高まったことを表す．

$$IMPL = \frac{\text{ミュータントの増加数}}{\text{論理 LOC}} \tag{3}$$

[*1] 「変換後」のコードがカンマ (,) で区切られている場合，そのミューテーション演算子が複数のミュータントを生成することを表す．

4. 実験

実験では，どのようなプログラム構造が SBFL に適するか明らかにする．

4.1 実験対象プログラムとテストスイート

実験対象プログラムは，3.1 節で示したデータセットに含まれる 728 個のメソッドのうち，以下の 4 つの条件を満たす 131 組のグループに含まれる 361 個のメソッドである．

条件 1： グループ内のメソッドが同機能である．
条件 2： グループ内のメソッドをテストスイートに通したときの命令網羅率と条件網羅率が 100% になる．
条件 3： グループ内のメソッド間で，プログラム構造が互いに異なる．
条件 4： グループ内のメソッド中に `switch` 文を含まない．

表 2: 本研究で新たに用いるミューテーション演算子

ミューテーション演算子	変換前	変換後
Change String Literal	`"String"`	`"String1"`
Change Instanceof	`a instanceof A`	`a instanceof B`
Nonvoid Method Calls	`a=method()`	`a=null`
Constructor Calls	`a=new A()`	`a=null`
Compound Operator	`a+=1`	`a-=1`
Change Numeric Literal	`if(x<0)`	`if(x<1)`
Change Boolean Literal	`true`	`false`
Change Unary Operator	`!ismethod()`	`ismethod()`
Add Not Operator	`if(b)`	`if(!b)`
More Specific If	`if(a&&b)`	`if(a)`, `if(b)`, `if(a||b)`
Less Specific If	`if(a||b)`	`if(a)`, `if(b)`, `if(a&&b)`
Break and Continue	`break;`	`continue;`
Null Assignment	`Object o1=o2;`	`Object o1=null;`
Empty Assignment	`s=a.toString();`	`s="";`
Primitive Assignment	`int a=b;`	`int a=0;`
Throw Exception	`throw new A;`	`throw new B;`

条件 2 から条件 4 を設定した理由を以下に示す．

条件 2

100%の命令網羅率は，対象プログラムのすべての文から疑惑値を得るために必要である．100%の条件網羅率は，様々な実行経路を持つテストケースを適用することで，各プログラム文の疑惑値をできるだけ異なる値にするために必要である．

条件 3

本研究の目的はどのようなプログラム構造が SBFL に適するかを調査することであるため，プログラム構造が同じ複数のメソッドを実験対象とする必要がないためである．

条件 4

メソッド中に `switch` 文が存在すると，正しく実行経路を取得できないためである．

条件の確認は以下の 2 つのステップで構成される．

(1) テストスイートを手動で修正

(2) 条件を満たさないメソッドを除外

以下，これらの手順の詳細を述べる．

Step1. テストスイートを手動で修正

各メソッドのテストスイートは Evosuite [12] によって自動生成された．自動生成されたテストスイートでは，条件網羅率が 100% にならない場合がある．よって，テストスイートを手動で修正し，条件網羅率が 100% になるようにした．

Step2. 条件を満たさないメソッドを除外

メソッドを目視で確認し，`switch` 文が含まれるメソッドを除外した．また，すべてのメソッドをテストスイートに通し，各メソッドが条件 1 から条件 3 を満たすかどうか確認した．その結果，合計で 145 組のグループと 367 個のメソッドが実験対象から除外された．

4.2 実験方法

実験は以下の 3 つのステップで構成される．

- **(A)** 実験対象プログラムの SBFL スコアを計測
- **(B)** SBFL スコアに差がないグループを考察対象から除外
- **(C)** SBFL スコアが高くなる要因を分類し，その理由を考察

以下では，これらの詳細を述べる．

Step A. 実験対象プログラムの SBFL スコアを計測

すべての実験対象プログラムについて，計測ツールを用いて SBFL スコアを計測する．計測ツールは SBFL スコアを計測するために，自動欠陥修正ツールである kGenProg [13] を用いて実行経路情報を取得する．kGenProg はプラグインとして JaCoCo[*2] を呼び出し，実行経路情報を得ている．各ミュータントに対して SBFL を実行した際の実行経路情報はファイルに保存され，計測後に確認できる．

[*2] `https://www.jacoco.org/jacoco/`

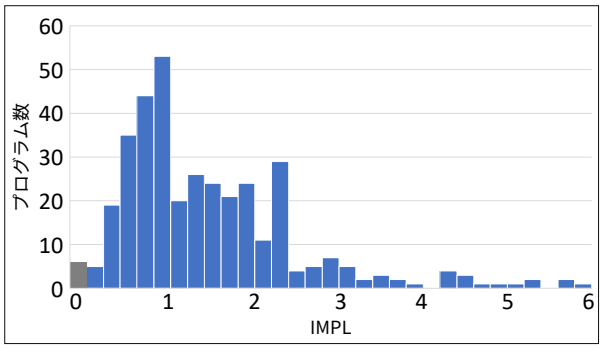


図 7: 各メソッドの IMPL の分布

また，計測ツール内のミュータント生成器に，2.5.2 節で示した 11 種類のミューテーション演算子と，3.2 節で示した 16 種類のミューテーション演算子を実装した．さらに，使用する演算子を切り替えられるようにした．これにより，演算子の追加前後におけるミュータント数の変化を記録できる．

Step B. SBFL スコアに差がないグループを考察対象から除外

プログラム構造が異なっても，SBFL スコアが同じになる場合が考えられる．SBFL スコアに差がないグループは，SBFL スコアが高くなる要因を考察するためには不必要であるため，Step C. の対象から除外する．

Step C. SBFL スコアが高くなる要因を分類し，その理由を考察

各グループに含まれるすべてのメソッドを目視し，グループをプログラム構造の差（要因）ごとに分類する．分類したそれぞれの要因について，統計的手法を用いて要因によって生じた SBFL スコアの差が有意であるかを検定する．また，メソッド間で SBFL スコアや実行経路情報を比較することにより，各要因によって SBFL スコアが高くなる理由を考察する．

5. 実験結果と考察

5.1 先行研究からの改善点の評価

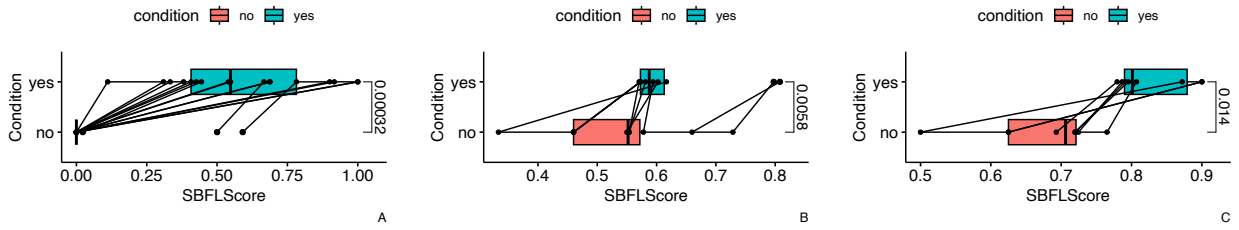
ミュータントの増加数

各メソッドについて IMPL を計測した結果を図 7 に示す．IMPL が 0 より大きいメソッドは 361 個中 355 個であった．よって，ほとんどのメソッドにおいて，SBFL スコアの信頼性は向上したといえる．

考察対象プログラムの増加数

実験対象である 131 組のグループのうち，119 組のグループで，グループ内のプログラムの SBFL スコアに差が見られた．よって，考察対象プログラム数は先行研究から大幅に増加したといえる．

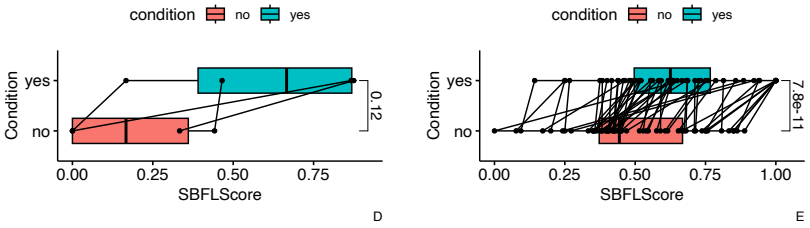
SBFL スコアに変化がないグループは 12 組存在した．変化が生じない理由は，以下の 2 つのいずれかであった．



(a) 三項演算子や Stream API 用いた文の代わりに制御文を導入している

(b) early return のための `if` 文を追加している

(c) `if` 文の条件式に論理演算子を用いず，別の文に分けている



(d) 最初の文で例外が発生しない

(e) 同一の条件分岐先で実行される文が少ない

図 8: 各要因の有無による SBFL スコアの差の分布

- グループ内のいずれのメソッドにも分岐が全く存在しない．
- グループ内のメソッドのプログラム構造の違いが，`if` 文に含まれるの複数の条件の出現順序だけである．

5.2 SBFL スコアが高くなる要因

SBFL スコアが高くなる要因として，新たに以下の 4 つが特定できた．

- 三項演算子や Stream API を用いた文の代わりに制御文を導入している
- early return [14] のための `if` 文を追加している
- `if` 文の条件式に論理演算子を用いず，別の文に分けている
- 最初の文で例外が発生しない

また，先行研究で発見された要因である「同一の条件分岐先で実行される文が少ない」に当てはまるグループも存在した．

さらに，ミューテーションの違いによって SBFL スコアに差が生じていると考えられるグループも存在した．SBFL スコアが高くなる要因と，それに当てはまるグループの数を表 3 に示す．

表 3: SBFL スコアが高くなる要因とグループ数

要因	グループ数
三項演算子や Stream API 用いた文の代わりに制御文を導入している	17
early return のための `if` 文を追加している	10
`if` 文の条件式に論理演算子を用いず，別の文に分けている	8
最初の文で例外が発生しない	4
同一の条件分岐先で実行される文が少ない	56
ミューテーションの違いによって SBFL スコアに差が生じている	9
不明	15

5.3 SBFL スコアの差の有意性

新たに発見した 4 つの要因と，先行研究によって発見された 1 つの要因を合わせた 5 つの各要因について，要因の有無による SBFL スコアの差の分布を図 8 に示す．また，SBFL スコアの差の有意性を確認するために，ウィルコクソンの符号順位検定を用いて p 値を算出した．図 8 から分かる通り，サンプル数が少ない「最初の文で例外が発生しない」以外の要因については，要因の有無による SBFL スコアの差が有意である ($p<0.05$) ことが確認できた．

5.4 SBFL スコアが高くなる理由

以下では，新たに発見した SBFL スコアが高くなる各要因について，なぜ SBFL スコアが高くなるかを具体例を用いて説明する．

三項演算子や Stream API を用いた文の代わりに制御文を導入している

図 9 のメソッド (b) には，`if` 文が用いられている．よって，テストスイートを実行した際の実行経路に差が生じるため，SBFL スコアが 0 よりも大きくなる．一方，メソッド (a) には，三項演算子が用いられており，制御文が存在しない．制御文がメソッド中にない場合，各文の疑惑値に

```
1 return a>b ? 1 : 0;
```

(a)

```
1 if(a>b)
2   return 1;
3 else
4   return 0;
```

(b)

図 9: `if` 文や `for` 文の有無を示すメソッドペア

```
1 int s = 0;
2 for (int i=0; i<l; i
      ++)
3   s += p[i];
4 return s;
```

(a)

```
1 if (1 < 1)
2   return 0;
3 int s = 0;
4 for (int i=0; i<l; i
      ++)
5   s += p[i];
6 return s;
```

(b)

図 10: `if` 文による early return の有無を示すメソッドペア

差が生じないため，各ミュータントの $rScore$ が 0 になり，SBFL スコアも 0 になる．よって，三項演算子や Stream API を用いた文の代わりに制御文を用いることで，SBFL スコアは向上する．

また，すでに制御文が存在するメソッドでも，三項演算子や Stream API を用いた文が存在する場合，それを制御文を用いて書き換えると SBFL スコアが向上する．

early return のための `if` 文を追加している

メソッドの入力によっては，長い処理を行う前に出力が分かる場合がある．例えば，図 10 のメソッド (a) では，配列の内容の処理に `for` 文が用いられている．一方，メソッド (b) では `if` 文を追加することにより，配列の長さが 0 である場合には `for` 文による処理を行わずに 0 を返す．

`if` 文による early return を追加することにより，一部の成功テストの実行経路が新たに追加した `if` 文に移動する．これにより，その後に続く文に欠陥がある場合に，その文を実行する成功テストの数が減少する．よって，欠陥がある文の疑惑値が上昇するため，SBFL スコアが向上する．

また，同様のことが例外発生にもいえる．例えば配列の添字として配列の長さ以上の値を与える場合のように，例外発生の条件が前もって分かる場合がある．その際に，early return と同様の方法で throw 文を追加することで，SBFL スコアが向上する．

`if` 文の条件式に論理演算子を用いず，別の文に分けている

図 11 のメソッド (a) では，条件式を論理 OR 演算子 (`||`) でつなぎ，1 つの `if` 文を用いている．一方，メソッド (b) では，条件式を分け，2 つの `if` 文を用いている．

条件式を論理 OR 演算子でつないだ場合，片方の条件式に欠陥がある場合でも，もう片方の条件式が正しい場合，その条件式をチェックするテストケースは成功テストとなる．よって，条件式を含む文が成功テストで多く実行される．成功テストで多く実行される文の疑惑値は低下するため，欠陥がない他の文の疑惑値の順位が上昇する．その結果，$rScore$ が低下し，SBFL スコアも低下する．反対に，条件式を別の `if` 文に分けると，SBFL スコアは向上する．

```
1 if (i < s || i > e)
2   return -1;
3 return 0;
```

(a)

```
1 if (i < s)
2   return -1;
3 else if (i > e)
4   return -1;
5 return 0;
```

(b)

図 11: 論理演算子の有無を示すメソッドペア

```
1 String target(String s, Object... args) {
2   StringBuilder b = new StringBuilder(s.
        length() + 16 * args.length);//s=
        nullで例外発生
3   int templateStart = 0;
4   ...
5 }
```

(a)

```
1 String target(String s, Object... args) {
2   s = String.valueOf(s);
3   StringBuilder b = new StringBuilder(s.
        length() + 16 * args.length);//s=
        nullで例外発生
4   int templateStart = 0;
5   ...
6 }
```

(b)

図 12: 例外発生位置の違いを示すメソッドペア

最初の文で例外が発生しない

本研究で用いた SBFL スコア測定ツールは，例外が発生した場合には，発生の原因となった文は実行されていないと判定する．これにより，SBFL スコアに影響が及ぶ場合がある．例として，図 12 に示す 2 つのメソッドを用いて説明する．図 12 のメソッド (a) では，入力として `s=null` が与えられた場合には 2 行目で例外が発生する．よって，例外発生を期待するテストケースを実行した際に，メソッド中のどの文も実行されていないと判定される．しかし，メソッド (b) では，入力として `s=null` が与えられた場合には 3 行目で例外が発生する．よって，例外発生を期待するテストケースを実行した際に，2 行目までが実行されたと判定される．3 行目以降に欠陥がある場合，`s=null` を与えると必ず 3 行目で例外が発生する．この場合，例外を期待するテストケースは成功テストになる．成功テストにより，実行された文の疑惑値が下がり，相対的に欠陥がある文の順位が向上する場合がある．これにより，最初の文

で例外が発生しないメソッド (b) の SBFL スコアが向上する．

6. 本研究の妥当性について留意する点

SBFL スコアの信頼性は，プログラム中で欠陥が生成できた箇所の多さによって決まる．よって，SBFL スコアの信頼性を測定するためには，ミュータントの数ではなく，ミューテーションを行った箇所がプログラム全体に分布しているかを見るほうが適切である．

SBFL の粒度としては，ステートメント単位以外にもブロック単位などが存在する [15]．本研究ではステートメント単位での SBFL を用いたが，他の粒度を用いた場合には実験結果が変化する可能性がある．

SBFL スコアの計測結果は，テストスイート及びミュータント生成器に影響を受ける．よって，異なるテストスイート及びミュータント生成器を使用すると，SBFL スコアの値が変化し，異なる結果が現れる可能性がある．

また，本研究とは異なるプログラムを実験対象とすると，今回の実験で得られた結果を否定する事例が現れる可能性がある．

本研究で用いた実行経路情報の取得ツールである JaCoCo は，例外発生の原因である文は実行されていないと判断する．実行経路情報の取得に別のプログラムを用いると，例外発生の原因である文が実行されたと判断され，実験結果が否定される可能性がある．

7. おわりに

本研究では，先行研究 [4] の課題であるミューテーション演算子の種類の少なさと実験対象プログラム数の少なさを解決した．また，SBFL スコアの計測実験を行い，SBFL 適合性が高いプログラム構造を調査した．調査の結果，新たに 4 つの SBFL 適合性が高いプログラム構造が得られた．

今後の課題としては，以下の 2 つが挙げられる．

自動変換ツールの作成

先行研究 [4] と本研究によって，SBFL 適合性が高いプログラム構造が複数得られた．よって，SBFL 適合性を高めるようにプログラムを自動変換するツールを作成することが可能であると考えられる．また，作成したツールを用いて，すでに欠陥が存在するプログラムを変換して SBFL を用いた欠陥限局の精度が向上するかを確かめる実験を行う．

SBFL 適合性と他の品質特性との関連性調査

SBFL 適合性が高くなる一方で，保守性が低下する場合がある．実際に，5.4 節で示したプログラムには，SBFL 適合性が高いプログラムにコードクローンが存在する．よって，SBFL 適合性と，保守性などの他の品質特性との関連性を調査することは重要な課題である．

謝辞 本研究は JSPS 科研費（JP20H04166, JP21K18302, JP21K11829, JP21H04877, JP22H03567, JP22K11985）の助成を得て行われた．

参考文献

[1] Hailpern, B. and Santhanam, P.: Software Debugging, Testing, and Verification, *IBM Systems Journal*, Vol. 41, No. 1, pp. 4–12 (2002).

[2] Wong, W. E., Gao, R., Li, Y., Abreu, R. and Wotawa, F.: A Survey on Software Fault Localization, *IEEE Transactions on Software Engineering*, Vol. 42, No. 8, pp. 707–740 (2016).

[3] Ali, S., Andrews, J. H., Dhandapani, T. and Wang, W.: Evaluating the Accuracy of Fault Localization Techniques, *Proc. International Conference on Automated Software Engineering*, pp. 76–87 (2009).

[4] 佐々木唯，肥後芳樹，柗本真佑，楠本真二：プログラムに対する欠陥限局の適合性計測，情報処理学会論文誌，Vol. 62, No. 4, pp. 1029–1038 (2021).

[5] Abreu, R., Zoeteweij, P. and Van Gemund, A. J.: An Evaluation of Similarity Coefficients for Software Fault Localization, *Proc. Pacific Rim International Symposium on Dependable Computing*, pp. 39–46 (2006).

[6] Abreu, R., Zoeteweij, P., Golsteijn, R. and van Gemund, A. J. C.: A Practical Evaluation of Spectrum-based Fault Localization, *Journal of Systems and Software*, Vol. 82, No. 11, pp. 1780–1792 (2009).

[7] Jia, Y. and Harman, M.: An Analysis and Survey of the Development of Mutation Testing, *IEEE TSE*, Vol. 37, No. 5, pp. 649–678 (2011).

[8] Martin, F.: *Refactoring: Improving the Design of Existing Code*, Addison-Wesley (1999).

[9] Coles, H., Laurent, T., Henard, C., Papadakis, M. and Ventresque, A.: PIT: A Practical Mutation Testing Tool for Java (Demo), *Proc. International Symposium on Software Testing and Analysis*, pp. 449–452 (2016).

[10] Xuan, J. and Monperrus, M.: Test case purification for improving fault localization, *Proc. International Symposium on Foundations of Software Engineering*, pp. 52–63 (2014).

[11] Higo, Y., Matsumoto, S., Kusumoto, S. and Yasuda, K.: Constructing Dataset of Functionally Equivalent Java Methods, *Proc. the International Conference on Mining Software Repositories*, pp. 682–686 (2022).

[12] Fraser, G. and Arcuri, A.: EvoSuite: automatic test suite generation for object-oriented software, *Proc. Symposium and the European Conference on Foundations of Software Engineering*, pp. 416–419 (2011).

[13] Higo, Y., Matsumoto, S., Arima, R., Tanikado, A., Naitou, K., Matsumoto, J., Tomida, Y. and Kusumoto, S.: kGenProg: A High-performance, High-extensibility and High-portability APR System, *Proc. the Asia-Pacific Software Engineering Conference*, pp. 697–698 (2018).

[14] Saca, M. A.: Refactoring improving the design of existing code, *Proc. IEEE Central America and Panama Convention*, pp. 1–3 (2017).

[15] Sarhan, Q. I. and Beszédes, c.: A Survey of Challenges in Spectrum-Based Software Fault Localization, *IEEE Access*, Vol. 10, pp. 10618–10639 (2022).