

## 単回帰分析の加重平均による Fault-Prone モジュール判別モデルの精度向上に関する研究\*

内垣聖史† 内田眞司† 門田暁人‡

†奈良工業高等専門学校 専攻科 電子情報工学専攻

‡奈良先端科学技術大学院大学 情報科学研究科

### 1 はじめに

ソフトウェア開発におけるテスト工程において、テストの効率化、およびソフトウェアの信頼性の確保に当たり、欠陥を含む確率の高い傾向にあるモジュール(以下, fault-prone モジュール)を特定することは重要である。そのため、複数のモジュールから計測されたソフトウェアメトリクス値を説明変数とし、バグの有無を目的変数とする fault-prone モジュール判別モデルが多数提案されている。

これまでに我々は、複数の単回帰モデルを寄与率により加重平均した fault-prone モジュール判別モデルを提案した[1]。提案モデルにおいて高い予測精度を確保できない理由の一つにメトリクス値の分布の違いがある[2]。具体的にはモデル構築に使用するプロジェクト(フィットデータ)と予測対象プロジェクト(テストデータ)間でのメトリクス値の分布の違いが影響していることが考えられる[3]。

本研究では、フィットデータとテストデータに前処理を行うことで提案モデルの予測精度の向上を目指す。本稿では、処理手法としてメトリクス値に対する標準化処理を行い2つのデータセット間の分布の違いを補正する手法を用いる。

評価実験として、12種類のプロジェクトデータを用いて、提案手法と従来手法のそれぞれで判別モデルを構築し fault-prone モジュールを予測した。その結果、12種類のプロジェクト中8種類のプロジェクトにおいて提案手法を用いたモデルの予測精度が向上した。

## 2 Fault-Prone モジュール判別モデル

### 2.1 概要

fault-prone モジュール判別モデル(以下, 判別モデル)とは、モジュールのメトリクス値を説明変数とし、目的変数であるモジュールに含まれるバグの有無を予測するためのモデルである。

判別モデルの構築は、過去に開発されたプロジェクトで計測されたデータセットを用いて行われる。構築した判別モデルに対して、新たに開発したモジュールのメトリクス値を入力することにより、モジュールに含まれる確率やバグの有無を予測する。

### 2.2 水野らのモデル[1]

水野らの提案モデルは、複数のサブモデルを寄与率により加重平均した判別モデルである。あるメトリクス $x_i$  ( $i=1, 2, \dots, n$ )の値を説明変数とし、バグの有無 $y$ を目的変数として、 $n$ 個のサブモデル $f(x_i)$ を構築する。判別モデルはサブモデルと寄与率 $w_i$ を用いて式(1)によって定義される。

$$y = \frac{\sum_{i=1}^n w_i f(x_i)}{\sum_{i=1}^n w_i} \quad \dots \text{式(1)}$$

本研究ではサブモデルに単ロジスティック回帰式を用いる。寄与率とは、サブモデルの目的変数に対する適合度をはかる指標であり、その値域は $[0, 1]$ である。寄与率が大きいほどサブモデルの適合度が高くなり、予測精度の向上が期待できる。

### 3 標準化処理

本稿では、データセットのメトリクス値に対する操作として標準化[3]を用いる。操作手順を以下に示す。

手順1. あるメトリクス値の各値 $X$ に1を加算し、底を10とした対数変換を行う。対数変換には式(2)を用いる。

$$Y = \log_{10} X \quad \dots \text{式(2)}$$

手順2. 各メトリクス値の平均値が0、標準偏差が1になるように正規化を行う。正規化には式(3)を用いる。ここで $\bar{Y}$ 、 $\sigma$ はそれぞれ $Y$ の平均値、標準偏差である。

$$Z = \frac{Y - \bar{Y}}{\sigma} \quad \dots \text{式(3)}$$

手順3. 手順1~2をデータセットの全メトリクス値に対して行う。

\* A study on the performance improvement of the Fault-Prone Module Detection Model by the weighted average of the Simple Regression Analysis

† Satoshi Uchigaki and Shinji Uchida,

Advanced Electronic and Information Engineering Course, Nara National College of Technology

‡ Akito Monden,

Graduate School of Information Science, Nara Institute of Science and Technology

## 4 評価実験

### 4.1 概要

本稿では、提案手法の有効性を測るために、提案手法と従来手法のそれぞれを用いて判別モデルを構築し、予測精度を比較した。評価実験の際、結果の妥当性を確保するため、実験を同一プロジェクトに対しそれぞれ 10 回繰り返した。また評価実験には、NASA IV&V Facility Metrics Data Program(以下、MDP) [4] で公開されている 12 種類のプロジェクトのデータセットを用いた。

### 4.2 評価方法

判別モデルの予測精度の評価方法として、Alberg Diagram の Area Under the Curve(以下、AUC)を用いた。AUC の値域は[0, 1]であり、予測精度の高いモデルほど AUC の値が大きくなり、ランダムに予測するモデルの場合、AUC の値は 0.5 程度となる。

### 4.3 実験手順

12 個のプロジェクトのデータセットに対して、以下の手順で実験を行った。

- 手順 1. データセットをランダムに 2 等分し、一方をフィットデータ (Fit)、他方をテストデータ (Test) とする。
- 手順 2. 提案手法のみ Fit, Test に対し標準化を行う。
- 手順 3. Fit に対して、判別モデルを構築する。
- 手順 4. 構築した判別モデルで、Test の fault-prone モジュールを予測する。
- 手順 5. 予測した結果を基に、Alberg Diagram を計算し、AUC を求める。

## 5 実験結果

評価実験より得られた各プロジェクトの AUC の平均値を表 1 に示す。12 種類のデータセットの内、8 種類のプロジェクト(表中\*で記す)で予測精度が向上した。全プロジェクトでは AUC が 0.037 向上した(有意水準 5%で有意差あり)。一方、標準偏差には有意な差が見られなかった。

最も AUC が向上したプロジェクト(PC4)のサブモデルの平均寄与率が標準化前と比べ 2 倍程度向上したのに対し、AUC が低下したプロジェクト(MW1)では平均寄与率の向上の割合が小さかった(表 2)。標準化によりサブモデルの寄与率が向上した事が、予測精度の向上につながったと考えられる。

## 6 まとめ

本研究では、水野らの判別モデルの予測精度向上を目的として、フィットデータとテストデータに標

表 1 各手法の AUC 平均値

プロジェクト名	従来手法	提案手法
*CM1	0.721	0.782
*JM1	0.621	0.667
KC1	0.748	0.747
*KC3	0.784	0.808
*MC1	0.865	0.910
MC2	0.636	0.622
MW1	0.770	0.740
*PC1	0.687	0.737
PC2	0.857	0.856
*PC3	0.734	0.796
*PC4	0.683	0.813
*PC5	0.871	0.943
平均	0.748	0.785

表 2 PC4, MW1 の平均寄与率

	PC4	MW1
従来手法	0.076	0.170
提案手法	0.158	0.181

準化処理を行う手法を提案した。

NASA が公開している 12 種類のデータセットを用いて提案手法と従来手法のそれぞれで fault-prone モジュール判別を行った結果、提案手法の予測精度が向上した。その理由として、標準化処理によりサブモデルの寄与率が向上した事が挙げられる。

今後は、異なるデータセット同士をそれぞれフィットデータ、テストデータとして実験を行い、結果の検討を行う予定である。

## 参考文献

- [1]水野, 内田, 門田, “バグモジュール率を従属変数とした Fault-Prone モジュール判別モデルの提案,” 情報処理学会第 70 回全国大会講演論文集, Vol. 1, pp. 323-324, March 2008.
- [2]Zimmermann T., Nagappan N., Giger E. and Murphy B.: Cross-project Defect Prediction. The 7th joint meeting of the European Software Engineering Conference and ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE'09), pp. 91-100 (2009)
- [3]藏本, 栢本, 亀井, 門田, 松本, メトリクス値の標準化による Fault-Prone モジュール判別モデルの精度向上, 情報処理学会研究報告, ソフトウェア工学研究会, October 2009.
- [4]NASA IV&V Facility Metrics Data Program, <<http://mdp.ivv.nasa.gov/>>