3L-3

WASRAモデルにおける欠損値処理法の違いによる予測精度の比較

松下 貴徳† 森居 諒‡ 内垣 聖史† 内田 眞司‡

奈良工業高等専門学校 専攻科 電子情報工学専攻† 奈良工業高等専門学校 情報工学科‡

1. はじめに

ソフトウェアテストおよび保守において，fault-prone モジュール（バグを含む確率の高いモジュール）を特定することは，テストの効率化，および信頼性を確保する上で重要である．

これまでに我々の研究グループでは複数の単回帰モデルを加重平均する WASRA モデルを提案し，従来手法より予測精度が向上することを確認した[1]．ただし，モデル構築時にはメトリクスに欠損値（未記録の値）がないことが前提となる．そのため，欠損値が存在する場合にはモデル構築前に何らかの方法で欠損値を補完，あるいは欠損値を含むモジュールの除去を行う必要がある．

本研究では，WASRA モデルにおける効果的な欠損値処理法を比較実験により明らかにする．

2. WASRA (Weighted Average of Simple Regression Analyses) モデル

WASRA モデルは，複数の単回帰によるサブモデル$f(x_i)$を寄与率w_iにより加重平均するモデルである．単回帰を用いているため多重共線性による予測精度低下の問題が生じない利点を持つ．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 個のサブモデルを構築する．判別モデルはサブモデルと寄与率を用いて式(1)によって定義される．

$$y = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i} \qquad (1)$$

本研究ではサブモデルに単ロジスティック回帰モデルを用いる．各サブモデルの重みづけに使用する寄与率とは，回帰式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回帰式の適合度が高くなり，予測精度の向上が期待できる．

Empirical Evaluation of Missing Data Techniques in WASRA model

† Takanori Matsushita and Satoshi Uchigaki
Advanced Electronic and Information Engineering Course, Nara National College of Technology

‡ Ryo Morii and Shinji Uchida
Information Engineering, Nara National College of Technology

3. 欠損メカニズム

本研究で用いた 3 種類の欠損メカニズム[2]を以下に説明する．

- **MCAR (Missing Completely At Random)**
  欠損値が生じる確率が，データ中のどの変数にも依存しない．
- **MAR (Missing At Random)**
  欠損値が生じる確率が，欠損する値以外の変数に依存する．
- **NM (Non-ignorable Missingness)**
  欠損値が生じる確率が，欠損する値そのものに依存する．

4. 欠損値処理法

比較をする 3 種類の欠損値処理法[3]を以下に説明する．

- **平均値挿入法**
  データセットの当該メトリクスの平均値を補完値とし，欠損値に挿入する．
- **リストワイズ除去法**
  欠損値を 1 つでも含むモジュールを全て除去する．
- **類似性に基づく補完法 (k-nn 法)**
  欠損値に対して，類似したモジュールのメトリクス値を用いて欠損値を補完する．

5 評価実験

5.1 実験概要

各欠損値処理法による予測精度を比較するために実際のソフトウェア開発で収集されたデータセットを用いて実験を行う．実験では，まずデータセットに対して各欠損メカニズムを適用し欠損値を発生させる．次に，欠損値発生後の各データセットに対して各欠損値処理法を適用し欠損値を補完，あるいは欠損値を含むモジュールの除去を行う．補完，モジュールの除去により欠損値を無くしたデータセットに対して fault-prone モジュール判別を行う．実験には NASA IV&V Facility Metrics Data Program で公開されている，欠損値の存在しない 6 種類のプロジェクトのデータセットを用いる．

5.2 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) を用いる．AUC の値域は[0,1]であり，予

測精度が高いモデルほど 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．

5.3 実験手順

6 種類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う．

手順 1. 各データセットに，MCAR, MAR, NM の 3 つの欠損メカニズムによって，欠損率 10%, 20%, 30%, 40%の欠損値を発生させる．

手順 2. 欠損値発生後のデータセットに，平均値挿入法，リストワイズ除去法，k-nn 法を適用して欠損値処理を行う．

手順 3. 各欠損値処法理適用後の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 4. Fit に対して判別モデルを構築し，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. 予測した結果により，Alberg Diagram を計算し，AUC を求める．

手順 6. 実験結果の信頼性を確保するために，手順 3 から手順 5 までを 10 回繰り返す．

6. 実験結果

AUC の平均値と標準偏差を欠損値処理法，欠損メカニズム，欠損率毎に表 1 に示す．各 AUC の値は全てのプロジェクトの全 10 回の平均値である．ただし，MCAR で欠損値を発生させたデータセットにリストワイズ除去法を適用した場合，モデル構築のための十分なモジュール数が残らなかったため，モデル構築ができなかった．

欠損メカニズムの違いによる予測精度には大きな差が見られない．また，全ての欠損値処理法において欠損値が含まれない場合が最も予測精度が高く，欠損率が高くなるに従って予測精度は低くなる．

各欠損値処理法を比較すると，平均値挿入法が最も予測精度が高く，リストワイズ除去法が最も予測精度が低い．リストワイズ除去法では，モデル構築のためのモジュール数が減ってしまうため，予測精度が低くなったと考えられる．

7. まとめ

本研究では，WASRA モデルにおける効果的な欠損値処理法を比較実験により明らかにした．

実験の結果，WASRA モデルに対しては，欠損値に対して平均値挿入法による補完が最も予測精度が高くなった．

今後の課題としては，今回実験に使用したデータセット以外でも同様の結果が得られるか，実験を行う必要がある．

参考文献

[1] 内垣ら, “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,” 情報処理学会第 73 回全国大会講演論文集, Vol.1, pp.433-434, March 2011.

[2] 柿元ら, “協調フィルタリングに基づく工数見積もり手法のデータの欠損に対するロバスト性の評価,” 電子情報通信学会論文誌 D, Vol.J89-D, No.12, pp.2602-2611, December 2006.

[3] 田村ら, “プロジェクト間の類似性に基づくソフトウェアメトリクスの欠損値の補完,”ソフトウェア信頼性研究会 第 4 回ワークショップ, pp.17-23, June 2007.

表 1　AUC の平均値と標準偏差

欠損値処理法	欠損メカニズム	欠損率									
		0%		10%		20%		30%		40%	
		平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
平均値挿入法	MCAR	0.754	0.086	0.748	0.085	0.746	0.084	0.736	0.086	0.739	0.082
	MAR			0.748	0.080	0.742	0.081	0.731	0.076	0.723	0.091
	NM			0.748	0.084	0.730	0.088	0.719	0.096	0.703	0.093
リストワイズ除去法	MCAR										
	MAR			0.707	0.104	0.696	0.094	0.689	0.110	0.683	0.103
	NM			0.675	0.110	0.672	0.086	0.668	0.110	0.658	0.162
k-nn 法	MCAR			0.716	0.085	0.695	0.081	0.664	0.070	0.648	0.058
	MAR			0.725	0.093	0.711	0.098	0.697	0.110	0.674	0.080
	NM			0.728	0.080	0.676	0.106	0.639	0.101	0.598	0.074