

# 原因分類情報を用いた Fault-Prone 予測モデルの適用事例

三菱総合研究所

仙頭洋一

sentto@mri.co.jp

## 開発における問題点

ソフトウェアリポジトリを用いたFPモジュール予測では、どういった修正原因であるかについての情報は十分ではなく、当該ファイル自体に欠陥が含まれている場合の修正や依存関係にあるファイルの修正に連動した場合の修正が一元的に扱われている。

## 手法・ツールの適用による解決

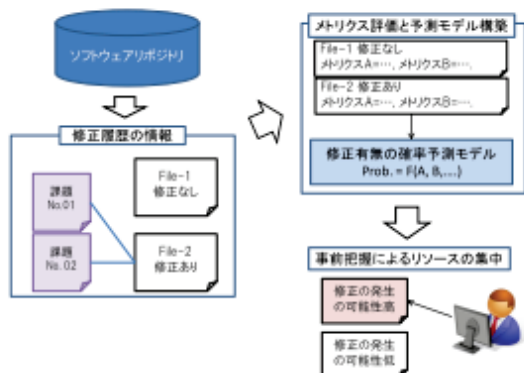
オープンソースソフトウェア開発におけるリポジトリ情報と開発者による修正原因分類情報を用いて、複数の定量的分析手法およびツール(R等)による分析に基づく変数選択を実践し、FPモジュール予測モデルの構築と、修正原因の違いによるFPモデル・説明力のある変数の違いを見出した。

## ヘッダ

### ①分析対象データ

→ Twitter4J

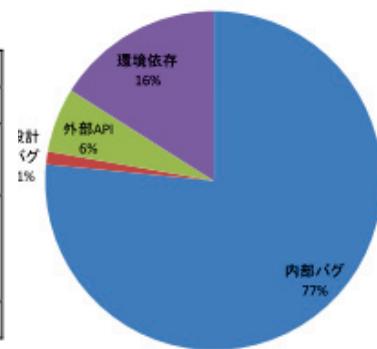
●開発者視点での原因分類情報を参考に修正データを集計



ソフトウェアリポジトリを活用したFP予測

## 修正原因分類情報

分類情報	内容
内部バグ	内部実装の問題に起因するもの
環境依存	特定の環境でのみ発生する問題に関わるもの
設計バグ	本来必要な引数がインターフェースに定義されていなかった、などのバグに起因するもの
外部 API	外部 API に起因するもの



題材としたTwitter4Jのバグ分類

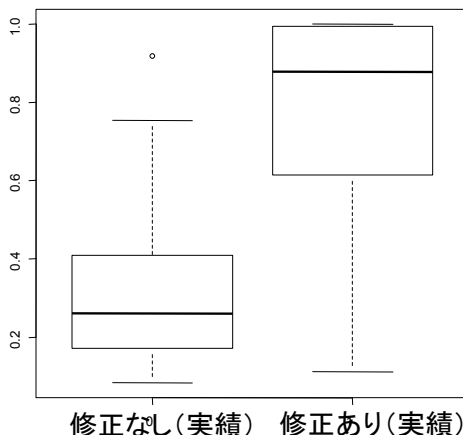
## 予測モデル構築と評価

### ②モデル化

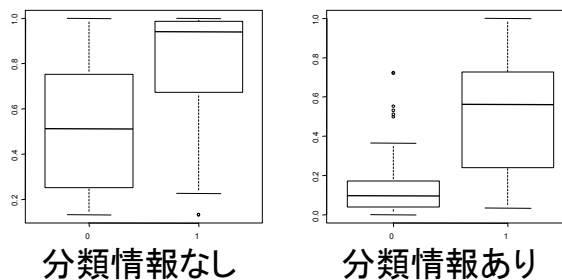
原因分類ごとに予測モデルを構築する。

- ロジスティック回帰モデルと複雑度メトリクス(後述)を用いる。
- モデルの説明変数として選ばれるメトリクスは原因分類によって違う

	(分類なし)		内部バグ		環境依存	
	係数	P値	係数	P値	係数	P値
定数	-2.11	****	-2.8	****	-2.09	***
LOCOM-HS	2.14	**	1.55		3.7	****
WMC	0.0746	**	0.147	**		**
DIT	-	*	0.879	*	-1.1	
NOC	-	-	-			
CBO	0.212	***	0.339	****		
RFC	-	-	-0.04	*	0.05	***
LOCOM-CK	-	-	0.00504		0.0033	1
NPM	-	-	-	-	0.196	**
AIC	112.54		111.28		93.07	



## 結果と課題



特定の分類に着目したモデルのほうが適合率、F値が高い

- 原因分類によって予測モデルは変化することを確認
  - 原因分類によって、説明力のある変数は異なる
  - 分類なしでの予測ができたとしても、取るべきアクションが違うので結局何もできない懸念がある
  - 予測結果を実際の現場での判断要素として使うためには、原因分類を踏まえた予測が必要