

11種類のfault密度予測モデルの 実証的評価

小林 寛武

戸田 航史

亀井 靖高

門田 暁人

峯 恒憲

鵜林 尚靖

60191008 大石 聖也

fault密度とは

- ・モジュールの中にあるバグの割合

例) fault密度の高いモジュール

→ 規模が小さくてバグが多いモジュール

fault-proneモジュールとは

- ・欠陥を含む確率の高いモジュール
- ・ソフトウェアテスト、保守工程において



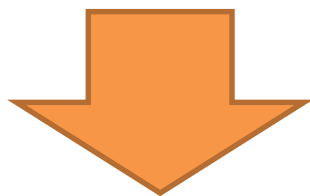
テスト工数の割り当て
レビューの優先順位の決定

従来研究

- 目的
限られたリソースで信頼性を確保する
- 提案された予測モデル
faultの有無、fault数、fault密度

従来研究の問題点(1)

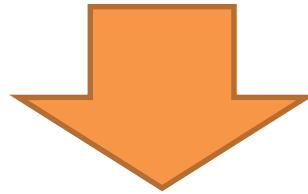
- ・性能比較を行う予測モデルの数が少ない
- ・性能評価に使うプロジェクト(データセット)の数が少ない



最も優れている予測モデルが論文ごとで異なる

従来研究の問題点(2)

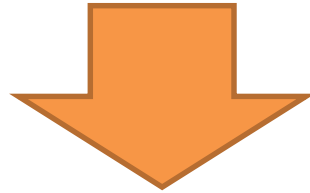
- fault密度予測モデルの性能評価、テスト工数を考慮したモデルの評価が行われていない



少ないテストでより多くのバグを発見する
→ fault密度の高いモジュールを予測する方が開発現場で役立つ

従来研究の問題点(3)

- ・モデル構築用と評価用のデータセットを同じサンプルから抽出している



実際の開発現場では、バージョンをまたがった予測モデルの評価が必要

本研究のResearch Question (RQ)

- RQ1

テスト工数を考慮した
fault密度予測モデルの評価

- RQ2

バージョンをまたがる
fault密度予測モデルの評価

RQ1の概要

- 11種類のfault予測モデルを比較
→ 予測性能の優劣を明らかにする

※11種類は予測モデルの6つのカテゴリーを網羅
ハイパパラメータ値の推定が行われている

ハイパパラメータとは

- ・モデル構築時に設定されるパラメータのうちモデル全体に影響を与えるパラメータ
- ・精度に大きな影響を与える

11種類の手法の概要

表1. 実験で使ったfault密度予測モデルとそのハイパパラメータ

カテゴリー	比較手法	概要	ハイパパラメータ
統計的アプローチ	LM	線形回帰	なし
	LARS	Least Angle Regression [25]	なし
	RVM	Relevance vector machine [26]	なし
近傍法	k-NN	k 近傍法	$k = \{1, \dots, 15\}$
	K*	エントロピーに基づく距離を使用した k 近傍法 [27]	なし
ニューラルネット	NNET	3 層ニューラルネット	$size = \{4, \dots, 28\}, decay = \{0.1, 0.2\}$
サポートベクターマシン	SVM	サポートベクターマシン	$C = \{2^{-6}, \dots, 2^{10}\}$
樹木モデル	M5	分類木	なし
	RPART	CART	$cp = \{0.05, 0.1, \dots, 0.7\}$
集団学習	GBM	ブースティングを用いた分類木	$n.tree = \{10, \dots, 50\}, shrinkage = \{0.1\}$
	RF	ランダムフォレスト	$mtry = \{10, 50, 100, 200, 250, 500, 1000\}$

RQ2の概要

- ・過去バージョンのデータ(メトリックス)を利用し
fault密度予測モデルを構築
- 次期バージョンに適用した場合の予測
性能を評価

対象データ

- ・25のOSSプロジェクトそれぞれから集められた各メトリックス値を記録したもの

※Git、CVSでバージョン管理を行い、
かつBugzilla、JIRAなどでバグ管理を行っているプロジェクトに限定

対象プロジェクトの 各バージョンについて

表2. 各バージョンでのモジュール数、fault数、ファイル数

プロジェクト名	略称	バージョン A		バージョン B							
		module	fault	file	module						
Clam AntiVirus	CLAM	1,597	93	8,723	56	OpenCms	OCMS	1,727	193	2,821	93
eCos	ECOS	630	110	3,480	67	openNMS	ONMS	1,203	123	1,416	112
Exim Internet Mailer	EXIM	184	28	61	31	OTRS	OTRS	420	111	763	169
freedesktop	FREE	1,982	477	2,855	603	PostgreSQL	POST	863	435	1,211	325
Ganymede	GANY	184	29	95	30	ProFTPD	PROF	195	64	220	61
Helma	HELM	312	38	100	17	Samba	SAMB	1,623	543	2,559	439
Hibernate Core	HIBE	406	87	5,032	1,191	Scilab	SCIL	2,636	239	1,248	244
HylaFAX	HYLA	137	70	109	35	Spring Security	SPRI	926	126	639	132
iptables	IPTA	151	96	276	91	squid	SQUI	250	59	185	92
KOffice	KOFF	930	363	990	232	Wine	WINE	1,627	584	1,382	288
NetBSD	NBSD	6,781	548	10,960	299	XDoclet	XDOC	105	30	135	5
OpenBSD	OBSD	1,706	275	5,964	158	XFree86	XFRE	713	350	761	259
						XORP	XORP	1,696	158	1,755	217

※Bの方がAよりも新しいバージョン

各RQで利用するデータ

- RQ1

Aからランダムに選んだ $2/3$ のモジュールを学習データ、残りの $1/3$ をテストデータ

- RQ2

Aを学習データ、Bをテストデータ

実験の評価に使う値

- faultのRecall値(再現率)
累積行数(調べた行数)がソースコード全体
行数の20%の時点で検出できる

→ 工数に対するfaultの発見効率

実験結果の表の見方

- AR
ある予測モデルの各データセットにおける
Recallの平均順位
→ 順位平均が低いほど、精度が高い
- AVG
各予測モデルのRecallの平均値

RQ1の実験結果

	LM	LARS	RVM	k-NN	K*	NNET	SVM	M5	RPART	GBM	RF
CLAM	89.5	91.1	32.3	34.7	90.3	16.9	66.9	90.3	67.7	70.2	94.4
ECOS	32.5	30.0	17.5	32.5	45.0	25.0	12.5	27.5	32.5	37.5	35.0
EXIM	25.0	25.0	0.0	12.5	50.0	37.5	50.0	0.0	0.0	25.0	25.0
FREE	43.1	48.8	24.7	54.1	54.4	34.1	41.3	56.3	8.4	52.8	54.4
GANY	12.5	12.5	8.3	25.0	16.7	25.0	25.0	25.0	0.0	29.2	33.3
HELM	21.7	21.7	8.7	30.4	26.1	13.0	13.0	21.7	0.0	30.4	21.7
HIBE	41.8	45.5	52.7	16.4	38.2	45.5	34.5	41.8	18.2	49.1	45.5
HYLA	29.9	63.5	17.5	21.2	52.6	21.9	15.3	35.8	35.0	24.1	65.7
IPTA	9.5	19.0	31.0	26.2	16.7	23.8	23.8	11.9	21.4	21.4	21.4
KOFF	23.1	28.3	25.9	32.1	32.1	30.2	23.1	34.0	34.0	28.8	35.4
NBSD	47.7	52.7	19.2	41.0	60.7	26.4	38.5	52.7	10.9	46.4	54.0
OBSD	53.9	60.7	13.5	44.4	59.6	42.1	37.6	63.5	6.2	48.9	60.1
OCMS	27.0	15.7	20.0	16.5	14.8	18.3	15.7	22.6	13.9	16.5	12.2
ONMS	26.9	19.2	26.9	30.8	53.8	46.2	17.3	40.4	15.4	46.2	36.5
OTRS	22.4	19.0	25.9	29.3	22.4	22.4	24.1	20.7	32.8	32.8	17.2
POST	16.1	42.2	25.8	40.1	40.7	41.0	35.9	42.6	15.2	41.3	39.2
PROF	35.9	33.3	9.0	23.1	28.2	6.4	26.9	25.6	25.6	14.1	23.1
SAMB	22.8	24.8	39.6	39.6	34.7	38.6	31.7	38.6	25.7	30.7	21.8
SCIL	39.0	42.9	35.2	41.0	51.4	40.0	26.7	58.1	31.4	58.1	56.2
SPRI	41.3	42.5	22.5	33.8	35.0	51.3	31.3	48.8	17.5	47.5	55.0
SQUI	26.2	29.5	9.8	21.3	14.8	16.4	14.8	16.4	6.6	27.9	27.9
WINE	30.6	50.6	22.2	35.7	39.5	24.4	33.0	51.7	16.2	32.2	35.3
XDOC	30.8	61.5	30.8	23.1	53.8	46.2	38.5	46.2	15.4	61.5	53.8
XFRE	22.8	45.7	34.9	41.9	43.3	42.2	40.8	45.3	37.0	41.2	47.4
XORP	58.6	58.6	25.3	54.0	69.0	34.5	31.0	71.3	56.3	56.3	50.6
AVG	33.2	39.4	23.2	32.0	41.7	30.8	30.0	39.5	21.7	38.8	40.9
AR	6.32	4.6	8.08	5.72	4.24	6.12	7.52	3.92	8.44	4.28	4.2

表3. RQ1: fault密度予測における各手法の性能(%)

RQ1の実験結果

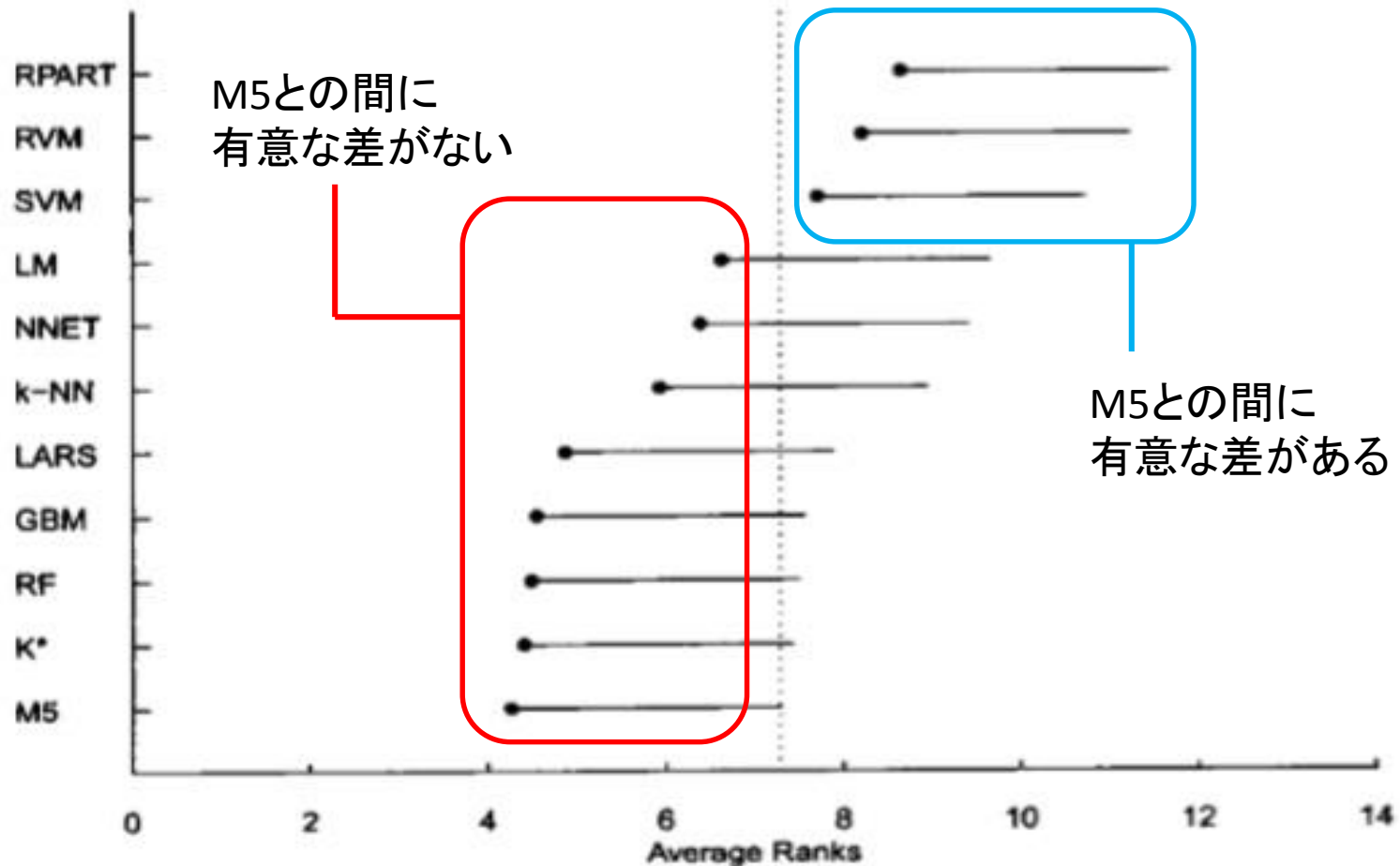


図1. RQ1でのDemsar's significance diagram

RQ2の実験結果

	LM	LARS	RVM	k-NN	K*	NNET	SVM	M5	RPART	GBM	RF
CLAM	76.1	93.6	41.3	43.1	82.6	15.6	66.1	55.0	53.2	97.2	93.6
ECOS	37.5	38.9	18.1	25.0	26.4	11.1	20.8	22.2	18.1	27.8	22.2
EXIM	20.5	20.5	17.9	25.6	7.7	53.8	12.8	10.3	10.3	23.1	15.4
FREE	41.2	50.0	31.8	59.8	58.2	60.2	40.1	60.2	22.7	59.9	59.8
GANY	25.5	17.6	21.6	21.6	13.7	23.5	9.8	19.6	15.7	11.8	17.6
HELM	8.3	8.3	25.0	8.3	12.5	37.5	16.7	8.3	16.7	8.3	4.2
HIBE	16.3	27.5	22.5	24.9	24.4	21.4	19.3	27.6	15.7	20.7	21.9
HYLA	35.0	38.0	27.0	25.0	29.0	28.0	36.0	46.0	34.0	36.0	34.0
IPTA	21.1	21.1	24.0	28.1	26.9	24.0	27.5	25.1	17.0	26.3	24.0
KOFF	25.2	23.9	19.5	28.1	28.1	30.8	23.5	29.5	28.3	22.5	29.3
NBSD	37.5	40.5	24.7	31.5	28.6	23.8	26.5	37.2	9.8	36.6	32.4
OBSD	47.1	51.6	24.0	35.1	40.4	29.3	15.1	51.6	12.9	30.7	39.6
OCMS	57.6	58.9	9.3	39.1	47.0	17.2	56.3	53.6	58.3	31.1	29.1
ONMS	26.8	24.8	22.9	24.2	33.3	21.6	15.7	27.5	20.3	24.8	31.4
OTRS	59.9	63.9	54.3	30.0	55.2	29.8	54.9	60.5	68.2	67.0	35.9
POST	24.8	24.6	30.2	29.6	32.0	32.2	28.2	34.5	23.6	27.9	29.4
PROF	38.1	24.4	20.8	11.9	14.3	22.6	24.4	27.4	31.5	26.8	24.4
SAMB	34.8	40.3	17.1	29.2	33.9	24.7	30.5	31.4	25.8	31.0	34.7
SCIL	26.9	26.9	21.8	24.6	32.1	23.8	28.0	33.2	19.7	33.4	31.6
SPRI	38.0	39.7	17.0	27.5	30.6	32.3	29.7	40.2	27.1	41.9	40.2
SQUI	25.4	38.9	25.2	24.9	26.4	19.7	31.2	35.2	4.5	28.4	21.4
WINE	35.2	41.3	16.2	38.4	33.1	27.2	34.0	45.2	21.7	43.9	37.8
XDOC	50.0	33.3	16.7	0.0	66.7	0.0	0.0	33.3	16.7	33.3	16.7
XFRE	31.7	29.6	30.4	35.0	40.9	32.4	25.6	39.8	33.3	31.5	33.3
XORP	49.2	54.1	16.9	42.9	50.0	36.6	38.3	63.9	48.1	40.2	50.3
AVG	35.6	37.3	23.8	28.5	35.0	27.2	28.4	36.7	26.1	34.5	32.4
AR	4.92	4.12	8.24	6.28	5.16	7.08	7.16	3.24	7.8	4.88	5.36

表4. RQ1: fault密度予測における各手法の性能(%)

RQ2の実験結果

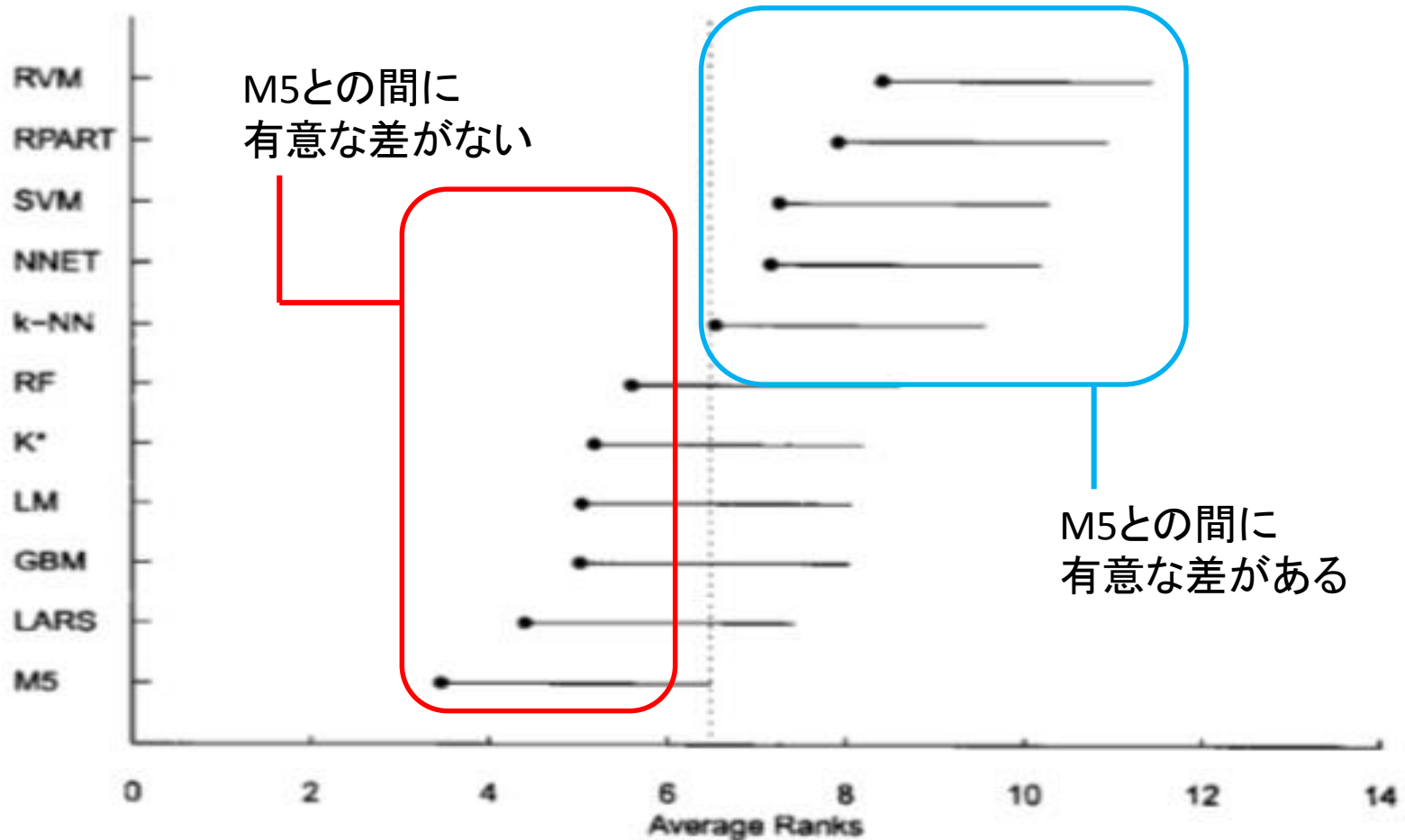


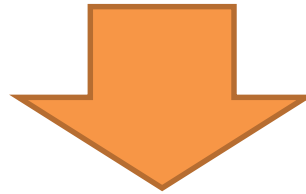
図2. RQ2でのDemsar's significance diagram

RQ1とRQ2の結果

- どちらでもM5が最も平均順位が低かった
→ M5が最も精度が高い
- M5が有意差を示す手法
RQ1では3種類、RQ2では5種類
→ バージョン間での予測において
他手法に比べてM5の精度が特に高くなる

まとめ

- RQ1、RQ2の結果を比較
→ ARが大きく変化した手法はなかった
- どちらともM5の性能が最も高かった



fault密度予測モデルにM5を使うことを推奨