

Revisiting the Impact of Classification Techniques on the Performance of Defect Prediction Models

Baljinder Ghotra, Shane McIntosh, Ahmed E. Hassan
Software Analysis and Intelligence Lab (SAIL)
School of Computing, Queen's University, Canada
{ghotra, mcintosh, ahmed}@cs.queensu.ca

International Conference on Software Engineering,
pp. 789–800, 2015

過去の研究結果(分類手法によって欠陥予測モデルの精度は変わらないとするもの)を改めて検証し, 欠陥予測モデルの精度は, 分類手法によって有意差があることを改めて実証.

背景

- 欠陥予測モデルの研究が佳境に入っている.
- 研究者は精度を高める分類手法を模索.
- 2008年に衝撃的な論文が発表される.
 - 「分類手法による予測モデルの性能の差はない」

“Benchmarking classification models for software defect prediction: A proposed framework and novel findings,”IEEE Transactions on Software Engineering,vol. 34,no. 4, pp. 485–496, 2008.
 - 17/22の分類手法は有意差がなかった...そうです.

本研究の目的とRQ

- 先行研究(Lessmannら)のアンチテーゼ
 - NASAのデータセット(noisy and biased)を使用
 - 1つのケーススタディしか行っていない
- 本研究で行うこと
 - 先行研究の再現
 - 先行研究のデータセットを正確にして再調査(RQ1)
 - OSSのデータセットでも有意差がないのか検証(RQ2)

データセットと調教方法

分類手法

Family	Technique	Abbreviation
Statistical Techniques	Naive Bayes	NB
	Simple Logistic	SL
Clustering Techniques	K-means	K-means
	Expectation Maximization	EM
Rule-Based Techniques	Repeated Incremental Pruning to Produce Error Reduction	Ripper
	Ripple Down Rules	Ridor
Neural Networks	Radial Basis Functions	RBFs
Nearest Neighbour	K-Nearest Neighbour	KNN
Support Vector Machines	Sequential Minimal Optimization	SMO
Decision Trees	J48	J48
	Logistic Model Tree using Logistic Regression	LMT
Ensemble Methods using LMT, NB, SL, SMO, and J48	Bagging	Bag+LMT, Bag+NB, Bag+SL, Bag+SMO and Bag+J48
	Adaboost	Ad+LMT, Ad+NB, Ad+SL, Ad+SMO and Ad+J48
	Rotation Forest	RF+LMT, RF+NB, RF+SL, RF+SMO, and RF+J48
	Random Subspace	Rsub+LMT, Rsub+NB, Rsub+SL, Rsub+SMO, and Rsub+J48

データセット

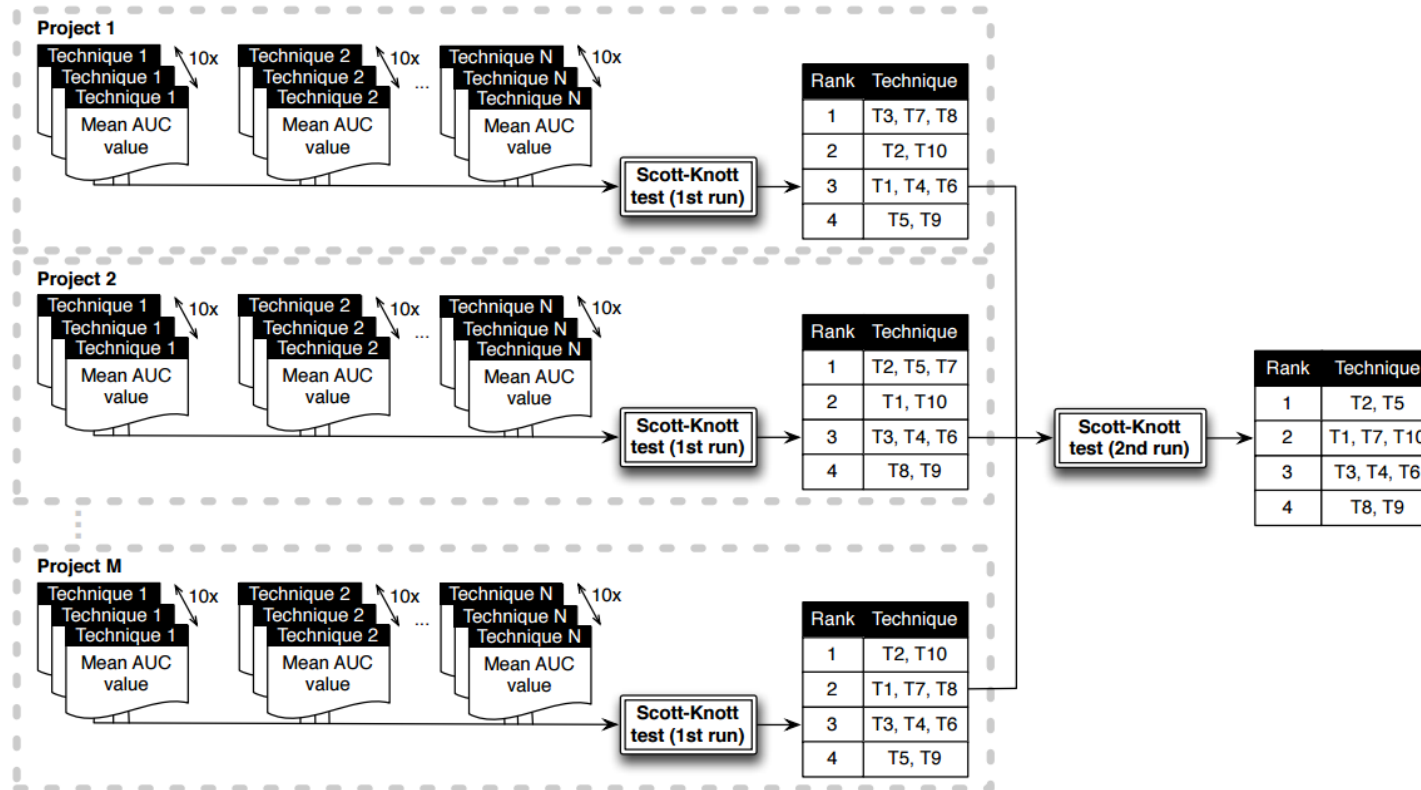
Corpus	Dataset	No. of Modules	Defective	Defective (%)
Noisy NASA (Section IV)	CM1	505	48	9.5
	JM1	10,878	2,102	19.3
	KC1	2,107	325	15.4
	KC3	458	43	9.3
	KC4	125	61	48.8
	MW1	403	31	7.6
	PC1	1,107	76	6.8
	PC2	5,589	23	0.4
	PC3	1,563	160	10.2
	PC4	1,458	178	12.2
Clean NASA (RQ1)	CM1	327	42	12.8
	JM1	7,720	1,612	20.8
	KC1	1,162	294	25.3
	KC3	194	36	18.5
	MW1	250	25	10
	PC1	679	55	8.1
	PC2	722	16	2.2
	PC3	1,053	130	12.3
PROMISE (RQ2)	Ant 1.7	745	166	22.3
	Camel 1.6	965	188	19.5
	Ivy 1.4	241	16	6.6
	Jedit 4	306	75	24.5
	Log4j 1	135	34	25.2
	Lucene 2.4	340	203	59.7
	Poi 3	442	281	63.6
	Tomcat 6	858	77	8.9
	Xalan 2.6	885	441	46.4
	Xerces 1.3	453	69	15.2

調教方法

10-fold-cross-validation

データを10個に分割し, 9で調教して, 1でテストをする手法

評価方法



double Scott-Knott test

各プロジェクト(データセット)でScott-Knottテストを行い、そこで得た順位をさらにScott-Knottテストを行うことで、評価する。

結果

- Lessmannらの再現
 - 分類できたランクは2つ
 - 確かに有意差なし

Overall Rank	Classification Technique	Median Rank	Average Rank	Standard Deviation
1	Ad+NB, EM, RBFs, Ad+SL, RF+NB, Rsub+NB, Ad+SMO, K-means, KNN, NB, SL, Bag+NB, Rsub+SL, Ad+J48, Rsub+J48, Ad+LMT, LMT, RF+SL, Bag+SL, RF+J48, Rsub+LMT, Bag+J48, RF+LMT and Bag+LMT	3.2	3.03	0.97
2	Rsub+SMO, SMO, RF+SMO, Ridor, Bag+SMO, Ripper and J48	8	7.91	1.04

- RQ1:綺麗なデータセットを使っても有意差はないか.
 - クリーンなデータセット

“Data quality: Some comments on the nasa software defect datasets,” IEEE Transactions on Software Engineering, no. 9, pp. 1208–1215, 2013
 - 4つのランクに分類できた
 - 分類手法による有意差を確認

- RQ2:PROMISEデータセット
 - 上と同じなので省略

Overall Rank	Classification Technique	Median Rank	Average Rank	Standard Deviation
1	Bag+J48, LMT, RF+J48, SL, Bag+SL, RF+SL, Rsub+SL, RF+LMT, Rsub+LMT, and Bag+LMT	1.22	1.24	0.15
2	RBFs, Ad+NB, KNN, Ad+SL, RF+NB, Rsub+NB, NB, Bag+NB, Ad+SMO, Ad+J48, Ad+LMT, and Rsub+J48	2.61	2.62	0.51
3	Ripper, EM, J48, and K-means	5.72	5.58	0.67
4	Rsub+SMO, RF+SMO, SMO, Ridor, and Bag+SMO	7.66	7.4	0.42

まとめと所感

まとめ

- 欠陥予測モデルの性能は分類手法による有意差がある
 - データセットに依存することを明らかにできた.
 - 様々な分類手法を試してみる必要がある

所感

- 先行研究のかなり詳細な調査が行われていた
 - 手法も確実に明快
- NASAのデータセット...
 - データをクリーンにする方法を調べたいと思った
“Data quality: Some comments on the nasa software defect datasets,” IEEE Transactions on Software Engineering, no. 9, pp. 1208–1215, 2013