

機械学習を用いたバグレポート分析研究の Bad Smells

目的

機械学習を用いたバグレポート分析に関する過去の研究を調査し、
悪い兆候(Bad Smells)を見つける

- ・ 2014年～2019年までの100件の論文を調査

発見されたBad Smells

1. 分析対象のプロジェクトが偏っている
2. 産業界で重要視されている問題が, アカデミアではあまり取り組まれていない
3. F1-scoreの算出方法
4. 仮設検定時に効果量を載せていない

Bad Smell 1 : 分析対象のプロジェクトが偏っている

- EclipseやMozillaは人気があり, 広く利用されているという理由で選定されがち
→ しかし, ソフトウェア工学やOSSにおいて代表的なプロジェクトというわけではない
- 実際には, これら特定プロジェクトのデータには偏りがあるため,
他のプロジェクトの代表として扱うのであれば, 正当な理由が必要
(多くの論文では正当な理由がなかった)

Bad Smell 2：産業界で重要視されている問題が、アカデミアではあまり取り組まれていない

調査方法

- ・アカデミアで取り組まれている研究テーマをカテゴリ化し, 論文本数の割合を計算
- ・実務者のインタビュー^[1]に基づき, 産業界で重視されている問題をランキング化

産業界とアカデミアで大きく差があったカテゴリ

カテゴリ	産業界	アカデミア
バグの重複検出	1位	1割以下
バグレポート改善, バグーコミット間のリンク付け	1位	ほぼ0%

所感

産業界(商用ソフトウェア開発)とアカデミア(OSS開発)では,
そもそも発生する問題が異なるので, 重要視される問題が異なるのは当然
あと比較するのに単位が違うのはヤバイ

[1] W. Zou, D. Lo, Z. Chen, X. Xia, Y. Feng, and B. Xu. 2018. How Practitioners Perceive Automated Bug Report Management Techniques. IEEE Transactions on Software Engineering (2018), 1-1.

Bad Smell 3 : F1-scoreの算出方法

交差検証で2値分類モデルを評価する場合

例として, 正例50件, 負例200件(一般的なデータは負例の方が多い)の4分割交差検証の結果をF1-scoreで評価

Type-II

Pr, Reそれぞれの平均値の調和平均

Fold	TP	FN	TN	FP	Pr	Re	F1-score
1	50	0	200	0	1.00	1.00	1.00
2	50	0	190	10	0.83	1.00	0.91
3	50	0	100	100	0.33	1.00	0.50
4	40	10	50	150	0.21	0.80	0.33

→ Precisionに偏りがあるモデル

Type-III

TP, FN, FPの合計値
を使ってF1-scoreを算出

Type-I

F1-scoreの平均値

Type-I	Type-II	Type-III
0.68	0.73	0.58

Type-I, IIは, 計算過程で平均を使うため, 平滑化される
Type-IIIの方が結果をよりよく反映できる