

Improving Bug Triaging with High Confidence Predictions at Ericsson

Aindrila Sarkar, Peter C. Rigby, Béla Bartalos

背景

- バグ修正は、開発及びメンテナンスにおいて不可欠
- バグを開発チームまたは開発者に割り当て
 - 手動だと時間がかかる
 - 割り当てを間違えると時間とコストが増加

目的

- バグを自動で正しいチームに割り当て
 - Ericssonでは、開発者でなく開発チームに割り当てている

RQ

1. Ericssonは既存のバグトリアーシのアプローチでうまく機能するか
2. アラームログとクラッシュダンプに含まれる情報はバグトリアーシに役立つか
3. トリアーシの精度に対する高い予測の影響は何か

使用するデータ

- ステータスがFIXED
- 2016年7月から2018年6月までに報告された
バグ
- 43チームが修正した11,570のバグレポート

データ属性

- テキスト
- カテゴリー
- ログ

TABLE II
BUG REPORT FEATURES USED IN THE MODELS

Type	Feature	Description
Text	Summary	A short description of the issue
Text	Description	Detailed description of the issue. <i>e.g.</i> , Configuration of the hardware, steps to reproduce the issue etc.
Text	Answer	Description of the final solution made to fix the issue
Categorical	Product	Product that the issue was found on
Categorical	Customer	Customer that uses the product the issue was found on
Categorical	Site	Location the issue was reported from <i>e.g.</i> , Ottawa-Canada, Kista-Sweden
Categorical	Submitter	Team that reported the issue
Categorical	Priority	Priority of the issue. <i>e.g.</i> , Major, Medium, Minor
Categorical	Configuration	Higher level category of the product. <i>e.g.</i> , Baseband, Radio Software etc.
Categorical	Generation	Generation of the product. <i>e.g.</i> , Gen1, Gen2 etc.
Log	Alarms	Software and hardware errors or warnings that can be potential threat for the system
Log	Crash Errors	Errors generated by a program or processor crash

データ属性(テキスト)

- 前処理
 - トークン化, ストップワード除去, ステミング処理
- 特徴抽出
 - TF-IDF[11]

$$\textbf{Definition 1. } TF-IDF_{t_i, b_j} = tf_{t_i b_j} \times \left(1 + \log \left(\frac{1+B}{1+b(t_i)} \right) \right)$$

[11]: A.-C. Florea, J. Anvik, and R. Andonie. Spark-based cluster implementation of a bug report assignment recommender system. In International Conference on Artificial Intelligence and Soft Computing, pages 31–42. Springer, 2017.

データ属性(カテゴリ)

- 特徴抽出
 - それぞれの特徴をバイナリ特徴ベクトル化

データ属性(アラーム)

- 処理
 - ログ処理ツールを使用してアラームを抽出
- 特徴抽出
 - LINE-IDF[2]

$$\textbf{Definition 2. } LINE-IDF_l = \left(1 + \log \left(\frac{1+A}{1+A_l} \right) \right)$$

[2]: A. Amar and P. Rigby. Mining historical test logs to predict bugs and localize faults in the test logs. In Proceedings of the 41st International Conference on Software Engineering. ACM, 2019.

データ属性(クラッシュダンプ)

- 処理
 - Ericssonツールを使用
 - ファイル名とエラーメッセージを抽出
- 特徴抽出
 - LINE-IDF[2]

$$\mathbf{Definition\ 2.}\quad LINE-IDF_l = \left(1 + \log\left(\frac{1+A}{1+A_l}\right)\right)$$

[2]: A. Amar and P. Rigby. Mining historical test logs to predict bugs and localize faults in the test logs. In Proceedings of the 41st International Conference on Software Engineering. ACM, 2019.

分類モデル&評価

- 分類モデル

- ロジスティック回帰

- 評価

- 時分割評価

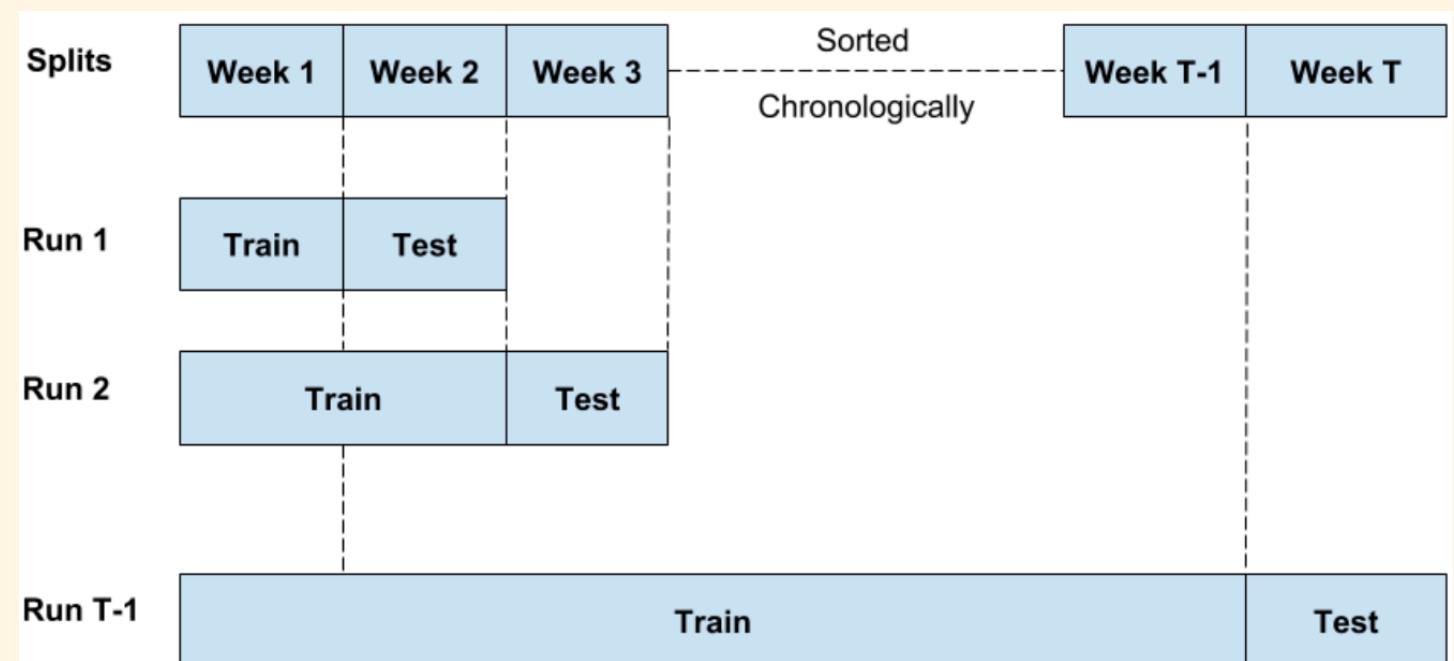


Fig. 5. Incremental evaluation setup

RQ(再掲)

1. Ericssonは既存のバグトリアーゼのアプローチでうまく機能するか
2. アラームログとクラッシュダンプに含まれる情報はバグトリアーゼに役立つか
3. トリアーゼの精度に対する高い予測の影響は何か

結果：RQ1&2

TABLE III
MODEL FEATURES AND PRECISION AND RECALL

Model	Features	Precision	Recall
M1	Text	62.49	64.13
M2	Categorical	73.70	74.00
M3	Text & Categorical	78.09	79.00
M4	Text, Categorical, Alarms & Crashes	77.96	78.85
M5	Alarms	15.57	22.04
M6	Crashes	11.52	22.21

- カテゴリは強力
- アラームとクラッシュダンプを追加しても精度は改善しない
- アラームはバグレポートの51.17%
- クラッシュログはバグレポートの5.52%

結果：RQ2(続き)

TABLE IV
MODELS OF BUG REPORTS CONTAINING ALARM LOGS

Model	Features	Precision	Recall
M7	Text & Categorical	70.72	72.80
M8	Text, Categorical & Alarms	70.09	72.28

- M7：アラームが含まれたバグレポートのみ使用
- M8：M7にアラームを追加
- アラームを追加しても精度は改善しない
- クラッシュダンプはレポートが不足
 - 同様の実験が行えない

RQ3(結果)

- 上位3チームの推薦
- 予測の信頼性が高い
バグレポートのみを使用

TABLE V
ACCURACY OF TOPN RECOMMENDATIONS

TopN	Accuracy
Top1	79.00
Top2	86.63
Top3	90.02

TABLE VI
TRIAGING BUGS WITH ABOVE A CONFIDENCE LEVEL CUTOFF. WITH A HIGHER CUTOFF, FEWER BUGS ARE TRIAGED, BUT THE ACCURACY OF THE PREDICTION IMPROVES.

Confidence Level	Triaged Bugs	Accuracy	Precision	Recall
≥ 0.1	100%	79.00	78.09	79.00
≥ 0.3	97.8%	80.19	79.36	80.19
≥ 0.5	89.18%	83.77	82.80	83.77
≥ 0.6	83.76%	85.73	84.59	85.73
≥ 0.7	78.56%	87.3	86.25	87.3
≥ 0.9	61.71%	90.17	89.73	90.17

トレーニング期間

- 古いデータを使用するとPrecision, Recallが下がる可能性がある[14], [27]

TABLE VII
RESULTS WITH LIMITED TRAINING DATA

Model	Features	Time Period	Precision	Recall
M3	Text & Categorical	2 months	73.67	75.66
M3	Text & Categorical	6 months	77.21	78.73
M3	Text & Categorical	All preceding data	78.09	79.00

- 古いデータを削除しても精度は向上しない

[14]: L. Jonsson, M. Borg, D. Broman, K. Sandahl, S. Eldh, and P. Runeson. Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts. *Empirical Software Engineering*, 21(4):1533– 1578, 2016.

[27]: X. Xia, D. Lo, Y. Ding, J. M. Al-Kofahi, T. N. Nguyen, and X. Wang. Improving automated bug triaging with specialized topic model. *IEEE Transactions on Software Engineering*, 43(3):272–297, 2017.

他の分類器

TABLE VIII
RESULTS FOR ALTERNATIVE CLASSIFIERS

Classifier	Precision	Recall
Naive Bayes	64.62	64.84
KNN	57.96	60.64
Linear SVM	77.74	78.08
Logistic Regression (M3)	78.09	79.00

- モデル3を使用
- ロジスティック回帰は
他の分類器よりも優れている

貢献

1. 時分割評価の場合, トリアーجزを実行するのには6ヶ月のデータで十分
2. 予測ではテキスト&カテゴリはログよりも優れている
3. 予測の信頼性が高いレポートを使用することで高い精度を達成