

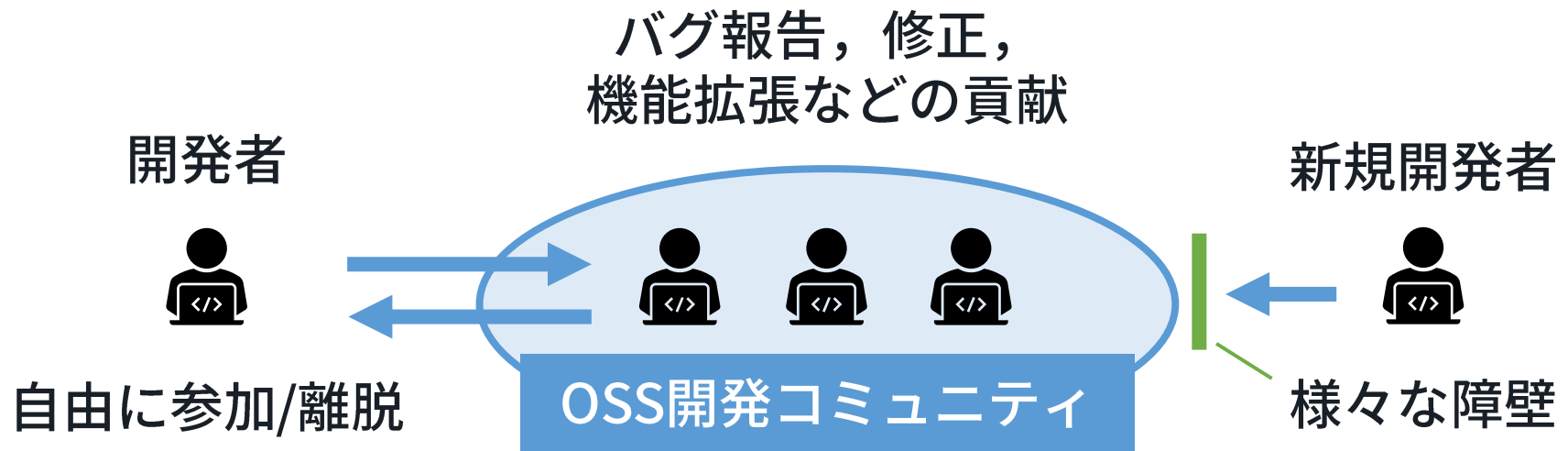
OSSプロジェクトへの オンボーディング支援のための Good First Issue自動分類

M2 堀口 日向

研究発表

2021/7/21

1. 背景



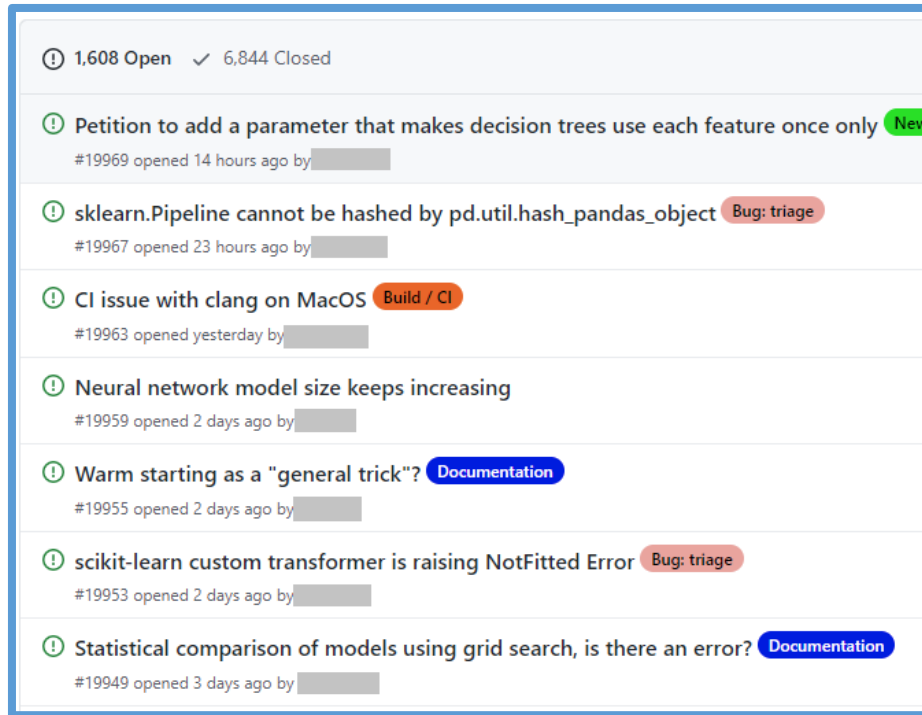
- プロジェクトを持続可能にするため,
新規開発者からの貢献を求めている^[1, 2]
- 新規開発者がコミュニティに参加する際,
様々な障壁に直面^[3, 4]

[1] Israr Qureshi and Yulin Fang. Socialization in Open Source Software Projects: A Growth Mixture Modeling Approach. Organizational Research Methods, pp. 208–238, 2011.
[2] Forte, A. and Lampe, C. Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature. American Behavioral Scientist 57, 5 2013, pp. 535–547.
[3] I. Steinmacher, T. Conte, M. A. Gerosa, and D. Redmiles, “Social barriers faced by newcomers placing their first contribution in open source software projects,” in CSCW ’15, 2015, pp. 1379–1392.
[4] C. Hannebauer and V. Gruhn, “On the relationship between newcomer motivations and contribution barriers in open source projects,” in OpenSym ’17, 2017, pp. 1–10.

1. 背景 | 課題選択の障壁

多くのOSSプロジェクトではIssue(課題)をもとに開発

Issue一覧の例 (Scikit-learn)



新規開発者

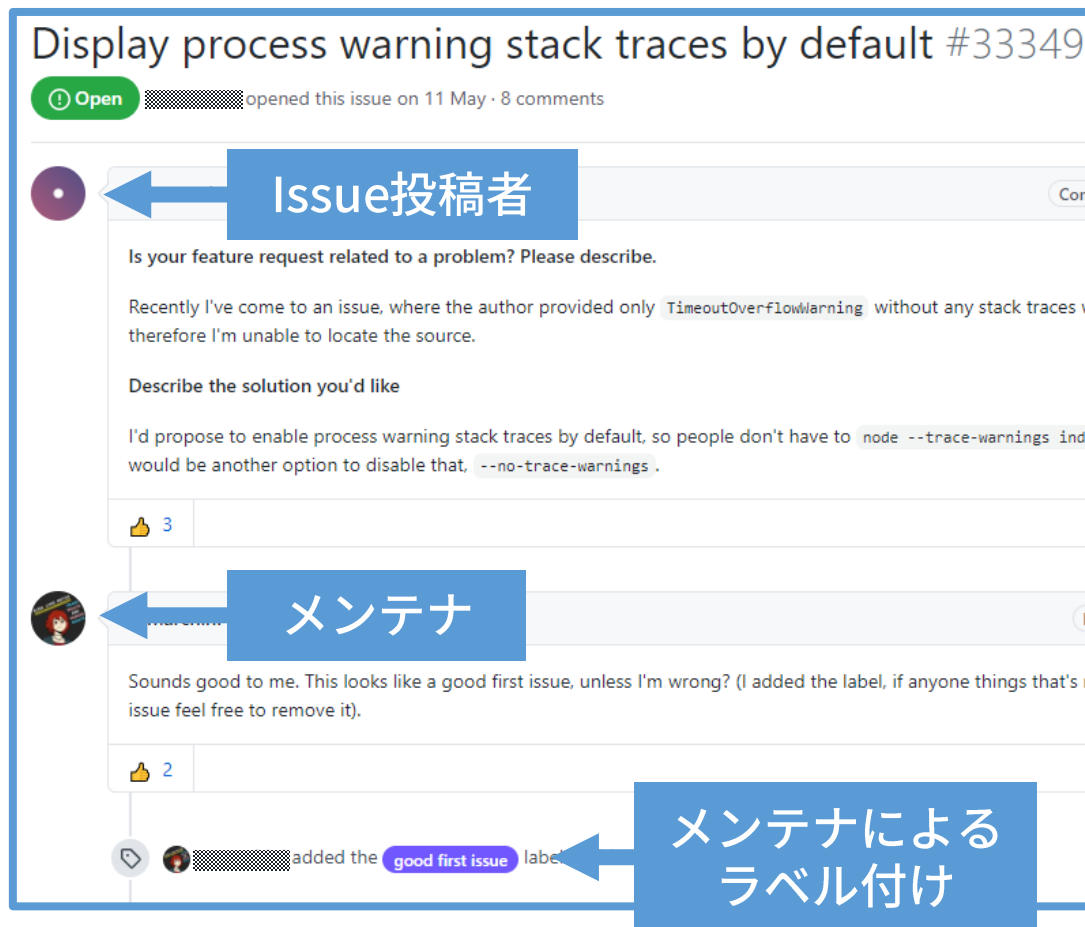


Issueの理解，対処に
時間がかかる^[5]

[5] I. Steinmacher, T. U. Conte and M. A. Gerosa, Understanding and Supporting the Choice of an Appropriate Task to Start with in Open Source Software Communities, 48th Hawaii International Conference on System Sciences, pp. 5299-5308, 2015

1. 背景 | Good First Issue (GFI)

GFIの例 (Node.js)



一部プロジェクトでは
新規開発者向けのIssueに
「**Good First Issue (GFI)**」
ラベルを付ける

メンテナにとって負担^[6]

[6] <https://github.blog/2020-01-22-how-we-built-good-first-issues/>

2. 目的とアプローチ

目的

新規開発者向けのIssueを分類する機械学習モデルを構築

アプローチ

GitHub上のGFIと通常のIssueを収集し、
ランダムフォレストを使って機械学習

リサーチクエスチョン

- RQ1: Issue投稿時に含まれる情報のみを利用してGFIを分類できるか？
- RQ2: 分類に重要なIssueの特徴量は何か？

3. 前回発表までの成果（1/2）

RQ1: Issue投稿時に含まれる情報のみを利用してGFIを分類できるか？

データセット

GFI数 > 500のリポジトリからIssueを収集

入力する特徴量

Issue投稿時に含まれる情報

- ・ タイトル，本文のBag-of-Wordsと単語数
- ・ 投稿者のリポジトリへの権限レベル(カテゴリ変数)

ランダムフォレストでGFI or 通常Issueの2値分類

3. 前回発表までの成果（2/2）

RQ1: Issue投稿時に含まれる情報のみを利用してGFIを分類できるか？

- 結果

評価指標	スコア
<i>Precision</i>	0.91
<i>Recall</i>	0.30

- Pros

GFIの中に通常Issueが紛れ込むことは少なく，
確実に新規開発者に適したIssueを分類可能

- Cons

GFIになりえるIssueを見逃しやすく，
新規開発者が貢献する機会が減る

4. RQ2: 分類に重要なIssueの特徴量は何か？（1/3）

重要度が高い特徴量上位10件

特徴量の名前	重要度
AuthorAssociation	0.0156
Body Words	0.0115
Title Words	0.0095
Body BoW-INLINE_CODE	0.0065
Body BoW-URL	0.0056
Body BoW-issue	0.0044
Body BoW-TIMESTAMP	0.0042
Body BoW-please	0.0039
Body BoW-VERSION	0.0037
Body BoW-add	0.0037

Issue投稿者の権限レベル
本文，タイトルの単語数

本文に含まれる単語
(大文字：前処理で置換)

4. RQ2: 分類に重要なIssueの特徴量は何か？（2/3）

Issue投稿者のリポジトリへの権限レベル

何らかの
権限を持つ
開発者

AuthorAssociation	GFI[件] (割合)	通常 Issue[件] (割合)
プロジェクトのオーナー	662 (0.05)	15 (0.00)
プロジェクトを所有する 組織のメンバー	3,662 (0.30)	21,873 (0.14)
プロジェクトの共同作業に 招待された開発者	772 (0.06)	2,488 (0.02)
過去にこのプロジェクトに コミットしたことがある開発者	4,560 (0.37)	61,392 (0.41)
上記のどれにも 当てはまらない開発者	2,727 (0.22)	65,227 (0.43)
合計	12,383	150,995

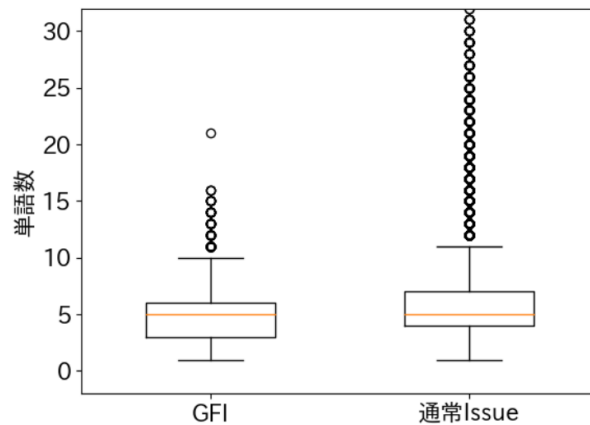
41%

16%

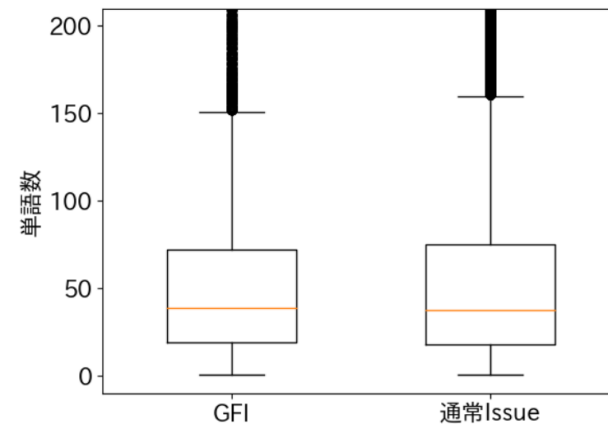
リポジトリに保守，運用に
関わる開発者による投稿が多い

4. RQ2: 分類に重要なIssueの特徴量は何か？（3/3）

タイトル，本文の単語数



タイトルの単語数



本文の単語数

- 分布に差はない
→ GFIを判別する特徴ではない？
- 外れ値が多い
→ 決定木のノード分割時に他クラスのデータが混ざりにくく
不純度が下がりやすいため，重要度が高くなる
外れ値を除去して学習したら，単語数の重要度が下がった

5. Recallが低い原因（1/2）

プロジェクトごとの分類結果

プロジェクト名	Precision	Recall
all	0.91	0.31
name-suggestion-index	0.81	0.96
material-ui	0.00	0.00
Cataclysm-DDA	1.00	0.03
zulip	0.00	0.00
radare2	0.00	0.00
pandas	1.00	0.03
game-of-life	1.00	1.00
tgstation	1.00	0.05
LeetCode-Solutions	1.00	0.99
elasticsearch	0.00	0.00
gutenberg	1.00	0.05
kibana	0.00	0.00
website	1.00	0.61
next.js	0.00	0.00
mattermost-server	0.93	0.80

全プロジェクトの結果

通常Issueがほとんどない
特殊なプロジェクト

3つのプロジェクトでしか
分類できていない

5. Recallが低い原因（2 / 2）

name-suggestion-indexのGFIを見てみると…

0 Open ✓ 120 Closed		Author ▼	Label ▼
🔒	Add Bodycare, Ramsdens, Klass, Neon Sheep and Greenhalgh's GB	add brand	good first issue multiple companies
#4290 by lakedistrictOSM was closed on 6 Oct 2020 5 of 5			
🔒	shop/convenience Конечка and shop/convenience Доброном	good first issue	
#4031 by krasnyekrylja was closed on 29 Jul 2020			
🔒	Add: Halfords (The Netherlands)	add brand	good first issue needs research needs wikidata
#3889 by boothym was closed on 11 Sep 2020			
🔒	Research needed for: shop/supermarket K+K	good first issue	needs research needs wikidata
#1700 by bhouse1 was closed on 9 Mar 2019			
🔒	Research needed for: shop/convenience 菜爾富	good first issue	
#1391 by bhouse1 was closed on 20 Jan 2019			
🔒	Research needed for: shop/convenience 全家便利商店	good first issue	
#1390 by bhouse1 was closed on 22 Jan 2019			
🔒	Research needed for: shop/convenience ローソンストア100	good first issue	
#1389 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience ローソン	good first issue	
#1388 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience ヤマザキショップ	good first issue	
#1387 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience ボブラ	good first issue	
#1386 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience セーブオン	good first issue	
#1385 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience セイコーマート	good first issue	
#1384 by bhouse1 was closed on 23 Oct 2018			
🔒	Research needed for: shop/convenience スリーエフ	good first issue	
#1383 by bhouse1 was closed on 16 Dec 2018			
🔒	Research needed for: shop/convenience サンクス	good first issue	
#1382 by bhouse1 was closed on 16 Dec 2018			

一般的なIssueから
GFIを分類するための
データセットしては不適切

6. 今後の予定

定型的ではないGFIを分類できるようにしたい

定型的なGFIをデータセットから外すとかはやるとして…

- そもそも分類性能が低すぎるのはなぜか？
 - GFIを見分ける特定の単語が存在しない？
文脈，類義語を考慮した特徴表現を使ってみる
- 学習時のテクニックの問題
 - 不均衡データ対策不足(現状クラスの重みづけのみ)
 - 過学習対策してない(ノードのサンプル数の下限設定など)
- 別のアイデア
通常Issueに取り組みづらい原因 = プロジェクトに不慣れ
→ そのIssueがどの程度プロジェクトへの理解を必要とするか
新規開発者にとって取り組みやすいIssueの指標になる