

Sustainability Forecasting for Apache Incubator Projects

Likang Yin, Zhuangzhi Chen, Qi Xuan, and Vladimir Filkov. 2021. Sustainability forecasting for Apache incubator projects. In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021). Association for Computing

紹介者：大森伊月

10月6日

背景

- 80%以上の OSS プロジェクトが時間の経過とともに放棄されており，特に小規模で若いプロジェクトが多い[44]
- OSSの持続可能性についての研究[7,9,49,53]は盛んに行われてきているが，時間的な予測を考慮されたものはない

[44] Charles M Schweik and Robert C English. 2012. Internet success: a study of open-source software commons. MIT Press.

[7] Narciso Cerpa, Matthew Bardeen, Barbara Kitchenham, and June Verner. 2010. Evaluating logistic regression models to estimate software project outcomes. Information and Software Technology 52, 9 (2010), 934-944.

[9] Jailton Coelho and Marco Tulio Valente. 2017. Why modern open source projects fail. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. 186-196.

[49] Didi Surian, Yuan Tian, David Lo, Hong Cheng, and Ee-Peng Lim. 2013. Predicting project outcome leveraging socio-technical network patterns. In 2013 17th European Conference on Software Maintenance and Reengineering. IEEE, 47-56.

[53] Marat Valiev, Bogdan Vasilescu, and James Herbsleb. 2018. Ecosystem-level determinants of sustained activity in open-source projects: A case study of the PyPI ecosystem. In Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 644-655.

ASF(Apache Software Foundation)

- ・ 開発者コミュニティが持続可能なOSSプロジェクトを構築・維持することを支援するため非営利団体
- ・ ASFに参加したいプロジェクトにサービスを提供するASF Incubator(ASFI)^[15]を運営している
- ・ ASFIのプロジェクトはASFのルールやサポートを受けてASFIを卒業することを目指す(持続可能性の成功)が、引退するプロジェクトも多い

STS(Socio-Technical System)

- ・ STSは個人的に交流を通じて知識を創造し共有する社会システムと、技術的なハードウェアを利用してタスクを達成する技術システムの2つのエンティティから構成される^[51]

[15] Juan C Dueñas, Félix Cuadrado, Manuel Santillán, José L Ruiz, et al. 2007. Apache and Eclipse: Comparing open source project incubators. IEEE software 24, 6 (2007), 90-98.

[51] Eric Trist. 1981. The evolution of socio-technical systems: A conceptual framework and an action research program. Ontario Ministry of Labour

目的とアプローチ

目的

- ASFIを卒業するプロジェクトとそうでないプロジェクトを区別する特徴を明らかにすること
- プロジェクト開発のどの時点でも持続可能性を予測できるモデルを構築すること
- プロジェクトの軌道修正のための実用的なアドバイスを提供すること

アプローチ

- ASFIが提供しているデータを用いて、各プロジェクトの時間的な社会的・技術的ネットワークを構築し、データ分析や予測モデルやアドバイスの実装する

社会技術ネットワークとモデル構築

- STSを用いて研究するために必要な社会的・技術的ネットワークを其々定義(例.開発者AからBへメールを送った場合, 社会的ネットワークの頂点AからBへ有向辺を生成)
- RNNの一種であるLSTM(長短期記憶)^[25]を用いて, あるプロジェクトのある月 m から $m+1$ 月の卒業予測を行うモデルを作成
- 解釈可能なモデルにするためにLIME^[41]を使用

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 1135–1144.

RQ

- RQ1

- 卒業したプロジェクトと引退したプロジェクトの間で、STSの測定値に有意な差があるか？

- RQ2

- ASFIプロジェクトの時間的なデータに基づいて、どの程度持続可能性を予測できるか？

- RQ3

- プロジェクトの持続可能性の状況を継続的にモニタリングできるか？

RQ1

卒業したプロジェクトと引退したプロジェクトの間で、STSの測定値に有意な差があるか？

- インキュベーション期間、コミット数、メール数、ソーシャルネットワークのノード数、テクニカルネットワークのノード数をASFIを卒業したプロジェクトと引退したプロジェクトで比較



Figure 2: The descriptive variables between graduated projects (in green, left) and retired projects (in red, right). The corresponding p-value of the Student's t-test is in the brackets, suggesting significant statistical differences exist between them.

卒業したプロジェクトと引退したプロジェクトの間には、複数の指標において有意な差が見られる

特に、引退したプロジェクトはインキュベーター期間に長く留まる傾向があり、卒業したプロジェクトの生産性と多様性は引退したプロジェクトに比べて高いことがわかった

RQ2

ASFI プロジェクトの時間的な痕跡に基づいて、どの程度持続可能性を予測できるか？

- ・ LSTMを用いた学習による予測とLIMEを用いたモデルによるどの変数により重点が置かれているかを確認

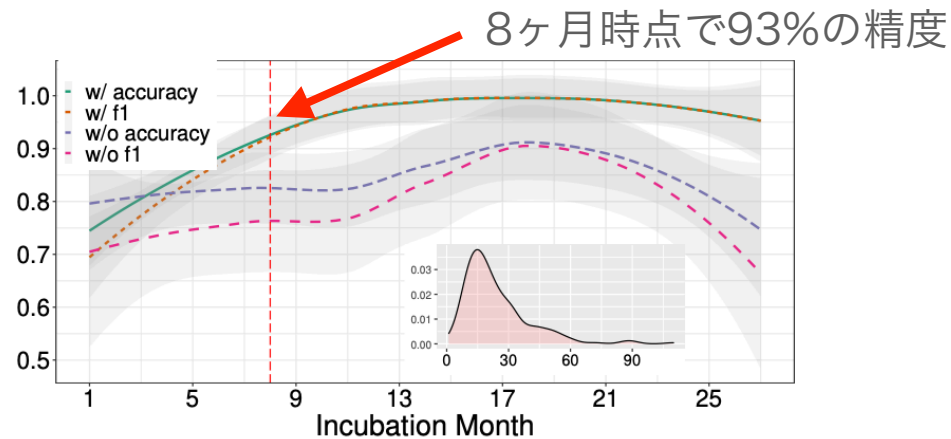


Figure 3: Performance metrics of the full LSTM model across incubation months (top 2 curves), showing the significant contribution of the socio-technical metrics. Curves are plotted using *loess*. Grey area shows the standard errors. The red vertical line shows 93% accuracy at 8 months of incubation. The inset shows project density over total incubation time.

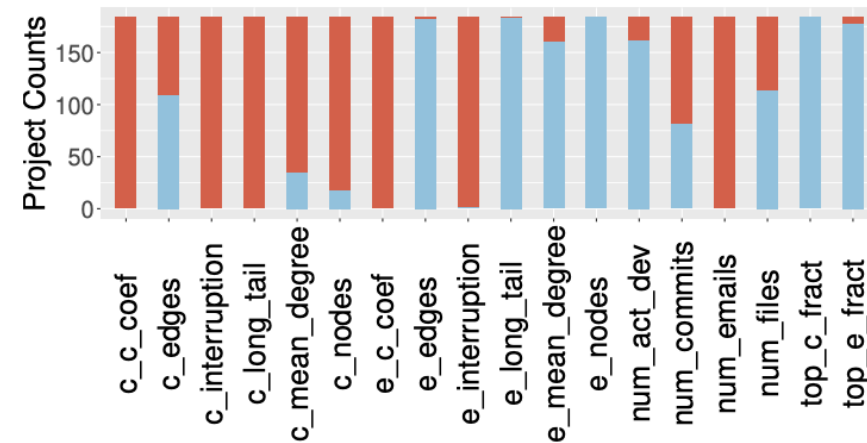


Figure 5: The overall-level coefficients of all variables of interest (blue is positive, while red is negative to graduation). It shows that some variables have same effect on almost all projects, while others do not.

プロジェクトの持続可能性に関する効果的なモデルは、複数の社会技術的な特徴とプロジェクトの特徴から構築することができる

安定した解釈可能なモデルが導出され、プロジェクトの特徴係数が得られた

RQ3

プロジェクトの持続可能性の状況を継続的にモニタリングできるか？

- ・ 卒業予測に大きな下降局面がないかどうかを継続的に監視し、検出された場合には予測を改善するための介入を提案
1. 大きな下降とは？
 - ・ 1～2ヶ月の間に予測値が5%以上低下した場合は、大幅な下降局面であると定義
 2. モデル変数と係数をどのように解釈して行動に移すか？
 - ・ 大幅に下降した月が検出された場合、プロジェクト開発者は、フィットしたモデルから離れた係数の特徴量を是正することを検討する

まとめ

- 目的
 - OSSプロジェクトの持続可能性を予測するモデルの構築
- アプローチ
 - STSを用いるための社会技術的ネットワークの構築
 - LIMEとLSTMを用いた機械学習
- 結果
 - プロジェクトの持続可能性を効果的にモデル化でき、プロジェクトの失敗と成功の理由を明らかにするとともに現状からの修正案を提案することができた

参考

- データセット
 - ASFに保存されているメーリングリストアーカイブ2つを使用して222の卒業・退学したプロジェクトのメールとコミットを収集
 - botによるコミットなどを除外する前処理を行い, 21328人の貢献者が判明

Table 1: Statistics of the 176 graduated and 46 retired projects in the ASFI dataset. c_* and e_* correspond to technical networks and social networks, respectively.

Statistic	Mean	St. Dev.	5%	95%
grad_status	0.79	0.41	0	1
num_files	1,821.87	3,346.23	122.45	5436.6
num_emails	3,963.12	4,930.54	262.65	12463.6
num_commits	12,451.84	27,373.41	453.8	36359.7
num_act_devs	121.23	119.85	25	415.05
c_interruption	0.20	0.20	0.03	0.60
e_interruption	0.11	0.14	0.01	0.32
top_c_fract	0.65	0.19	0.38	0.94
top_e_fract	0.71	0.11	0.49	0.85
c_nodes	15.44	17.09	2	49.9
c_edges	120.15	276.19	1	531.5
c_c_coef	0.78	0.25	0	1
c_long_tail	10.65	11.93	0	33.85
c_mean_degree	8.14	7.45	1	23.75
e_nodes	113.38	115.07	22	408.15
e_edges	399.23	562.38	47.1	1315.9
e_c_coef	0.43	0.10	0.28	0.58
e_long_tail	10.33	7.15	3	24.95
e_mean_degree	6.07	1.91	3.88	9.62

参考2

- LSTMに用いた変数
(接頭語のc_はcode,
e_はemailを示す)

変数名	説明
num_files	ユニークなファイル数
num_emails	メール数
num_commits	コミット数
num_act_devs	アクティブに活動した貢献者数
c_interruption	コミットがなかった最大期間
e_interruption	メールがなかった最大期間
top_c_fract	上位10%の貢献者のコミット割合
top_e_fract	上位10%の貢献者のメール割合
c_nodes	技術的ネットワークのノード数
c_edges	技術的ネットワークのエッジ数
c_c_coef	技術的ネットワークのクラスタリング係数
c_long_tail	技術的ネットワークの第3四分位数
c_mean_degree	技術的ネットワークの平均度数
e_nodes	社会的ネットワークのノード数
e_edges	社会的ネットワークのエッジ数
e_c_coef	社会的ネットワークのクラスタリング係数
e_long_tail	社会的ネットワークの第3四分位数
e_mean_degree	社会的ネットワークの平均度数