Analyzing the Factors Impacting Open-Source Project Aliveness

Rudi Chen and Ivens Portugal

Outline

- Introduction
- Approach
- Study Design
  - Features
  - Labels
- Data Cleaning
- Results
- Conclusions

Introduction

Who wants to know if a project will be alive in the future?

- Developers that contribute
  - Whether their contributions will last and impact other developers
- Developers who use the source code/library
  - Whether project has updated security patches, dependencies, compatibility with new technologies

Approach

- Investigate different factors and their influence on open-source project aliveness prediction
  - RQ1: What **contribution activity** impacts open-source project aliveness prediction?
  - RQ2: What is the impact of the **founder** on open-source project aliveness prediction?
  - RQ3: How does **documentation quantity and activity** affect open-source project aliveness?

Study Design

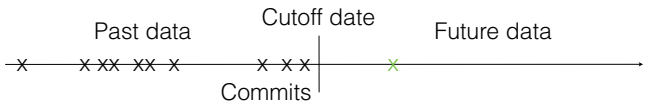
- Boa infrastructure and language
  - State University of Iowa
  - 8 million projects
  - Compiled MapReduce jobs
  - Distributed processing
- Use a Decision Tree algorithm and statistical tools to observe the features that best predict project aliveness

Features

- Number of committers
- Number of committers with multiple commits
- Number of committers sticking at least a day
- Percentage of commits - founder
- Days after last commit
- Density of commits
- ...

Labels and backtesting

- Project Aliveness
  - A project is said to be alive when there is at least one commit after the observation date.
- Backtesting
  - Choose a cutoff date.
  - Data before the date is used for the feature. Data after the date is used for the labels.
  - Repeat.



Data Cleaning

- Data before September 2013
- Only **Java** projects had commit history information
- Single-owned projects were discarded - not collaborative
- Projects with commit activity restricted to a week were discarded - student projects, Hackathon projects, weekend experiments, etc
- 68,299 projects were considered (20%)

Results (RQ1 - Features)

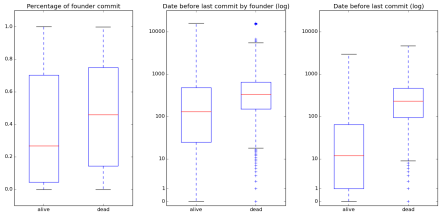
- Features with more impact
  - **Date** since last commit - 80.72% decision tree accuracy
  - **Date** since last founder commit
  - Commit density
  - Largest number of days between two consecutive commits

```
date_before_last_commit <= 37
|   date_before_last_commit <= 6: alive
|   date_before_last_commit > 6
|   |   project_age <= 154
|   |   |   date_before_last_commit <= 17: alive
|   |   |   date_before_last_commit > 17: dead
|   |   project_age > 154: alive
date_before_last_commit > 37
|   date_before_last_commit <= 134
|   |   project_age <= 223: dead
|   |   project_age > 223
|   |   |   date_before_last_commit <= 85: alive
|   |   |   date_before_last_commit > 85
|   |   |   |   project_age <= 546: dead
|   |   |   |   project_age > 546: alive
|   date_before_last_commit > 134: dead
```

```
density_of_commit <= 0.211111
|   largest_gap_between_consecutive_commits <= 43: dead
|   largest_gap_between_consecutive_commits > 43
|   |   percentage_top5_contributors_commit <= 0.995775: alive
|   |   percentage_top5_contributors_commit > 0.995775
|   |   |   density_of_commits <= 0.068548: dead
|   |   |   density_of_commits > 0.068548: alive
density_of_commits > 0.211111
|   largest_gap_between_consecutive_commits <= 18
|   |   density_of_commits <= 0.939239: dead
|   |   density_of_commits > 0.939239: alive
|   largest_gap_between_consecutive_commits > 18: alive
```


Results (RQ2 - Founder)

- Impact of Project Founder
  - Lower percentage of project founder commit means higher chances of project alive in the future
  - A community has been created around the project, contributing to its maintenance and aliveness



Results (RQ3 - Documentation)

	Last commit within 7 days			Last commit within 30 days		
	Alive	Total	%	Alive	Total	%
readme	2147	41782	**5.1**	5252	41782	**12.6**
license	1035	16210	6.3	2392	16210	14.8
todo	582	8943	6.5	1256	8943	14.0
install	359	4855	7.4	680	4855	14.0
contributing	218	3028	**7.2**	448	3028	**14.8**
changelog	243	3407	7.1	498	3407	14.6
One of the above	2835	53904	5.3	6752	53904	12.5
Two of the above	1203	16641	7.2	2641	16641	15.9
All projects	**3394**	**68299**	**5.0**	**22377**	**23302**	**11.8**

Analyzing the Factors Impacting Open-Source Project Aliveness

Rudi Chen and Ivens Portugal