NTDS-2016 Project

Pavlos Nikolopoulos, Matthaios Olma, Stefanos Skalistis

December 5, 2016

The idea

A discouraging factor for adopting open source software (OSS), either for academical or industrial purposes, is the uncertainty of whether an OSS project will continue to have support by the community until the end of its use. For this reason, the common practice followed by companies, is to use only well established and supported OSS projects.

In this work, our goal is to provide a way to predict whether an OSS project will remain active and supported for a given period of time. We believe that such prediction is not trivial, since there are several factors contributing to the success of a project such as human dynamics, project acceptance by the community and usefulness. This problem becomes harder, when considering that these factors change over time as the users and developers of OSS projects interact and affect each other.

Project Plan

To predict whether a project will remain active for a given period of time T (e.g., 6 months), we will proceed as follows:

- Extract metadata regarding OSS projects from a open source repository (e.g., github, sourceforge)
- Choose a subset of metadata that summarizes the activity of each project. A candidate set of metadata consists of: number of active users, number of developers, rate of commits, rate of downloads, rate of wiki posts, rate of stackoverflow posts, rate of Facebook/Tweeter posts. The final dataset will have the format of a multidimensional time series containing the set of metadata for each point in time.
- Label the projects as *successful* or *unsuccessful*, according to whether they are still supported after time T, i.e., there is active bug fixing or feature development.
- Separate the dataset into two sets, training and validation.
- Create and train a NN that should be able to predict the success of a newly created OSS project i.e., a project that exists for time $t < T$ (e.g., a month or so). For this step, we will deal with all the challenges included in designing an NN, such as defining its depth, width, weight matrix initialization, regularization and optimization methods, so that we maximize the accuracy of the prediction.

Therefore, the input to the neural network will be a set of time series representing the metadata of the OSS projects over time t and the output will be the likelihood of the project being successful.