POLITECNICO DI TORINO

Master degree course in Computer Engineering

Master Degree Thesis

Time prediction of software development via machine learning

Artificial Intelligence applied to Software Engineering

Supervisors
prof. Maurizio Morisio

Candidates
Jacopo Nasi [255320]

Internship Tutor
dott. Davide Piagneri

Anno accademico 2019-2020

Summary

La pressione barometrica di Giove viene misurata mediante un metodo originale messo a punto dai candidati, che si basa sul rilevamento telescopico della pressione.

Acknowledgements

Un ringraziamento speciale ai cavalieri di Smirnuff, luce della mia battaglia.

Contents

Chapter 1

Introduction

1.1 General Problem

Forecasting is one of the most critic part of a company, it could drive to easily success as well as drive to failure. A software project is not different from a manufacturing product, its development, infact, require analysis of different kind, from resources needed to costs and time required.
The software development experience shows that the process of analysis is really difficult, due to the nature of the problem, coding is a mind product and the time required to produce it can varying in accord to a lot of different factors.

1.2 Tools used

This work is mainly conducted using software tools, here a list of the tools used:

Python The main programming language of the thesis project. Used for data management, feature extraction, machine learning models and for interfaction with other softwares. The specific version used is the v3.7.0

Pandas Open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

NumPy Scientific computing with Python.

Matplotlib 2D Plotting library for Python.

Seaborn Another plotting library for Python.

Tensorflow Platform for machine learning.

Keras High level API for neural networks.

SciKit-Learn Tools and libraries for machine learning.

GitLab Sourcing platform based on Git. Used for the code of the project, available here:
https://gitlab.com/EiS-Projects/analytics/temp/thesisProjectJN.

GitHub Sourcing platform based on Git. Used for the thesis and calendar sourcing:

- Thesis: https://github.com/Jacopx/Thesis
- Calendar: https://github.com/Jacopx/ThesisCalendar

JetBrains IDEs Student-free IDE for different language development, product used:

- PyCharm: https://www.jetbrains.com/pycharm/
- DataGrip: https://www.jetbrains.com/datagrip/

Chapter 2

State of art

2.1 Literature

Speaking about other works.

Chapter 3

Datasets

The following section illustrate the structure of the all the principal datasets used during this thesis project.

3.1 SEOSS33

The SEOSS33[1] is a dataset collecting bug, issue, reports, commit and lot of other information of 33 open source project, following their progress via sourcing platform. The dataset is enriched also with time stamps, release versions, component information and developer comments.
At today there are no other public research conducted over this datasets, this works seems to be first.
The data is retrived by the issue tracking system (ITS) and the version control system (VCS), the process is shown in figure 3.1.

To unify the project specific difference, the typed issue, e.g. *New Feature* or *Bug Report*, are mapped to five issue categories:

- Bug: A problem which impairs or prevents the functions of the product
- Feature: A new feature of the product
- Improvement: An enhancement to an existing feature
- Task: A task that needs to be done
- Other: Various

The study collects 33 projects that are using Atalassian Jira and git, for popularity reasons. Is also required that the projects in the dataset should be majorly written in one programming language, Java is choosen. Due to machine learning nature, the choosen project must have great number of issue, all project were in

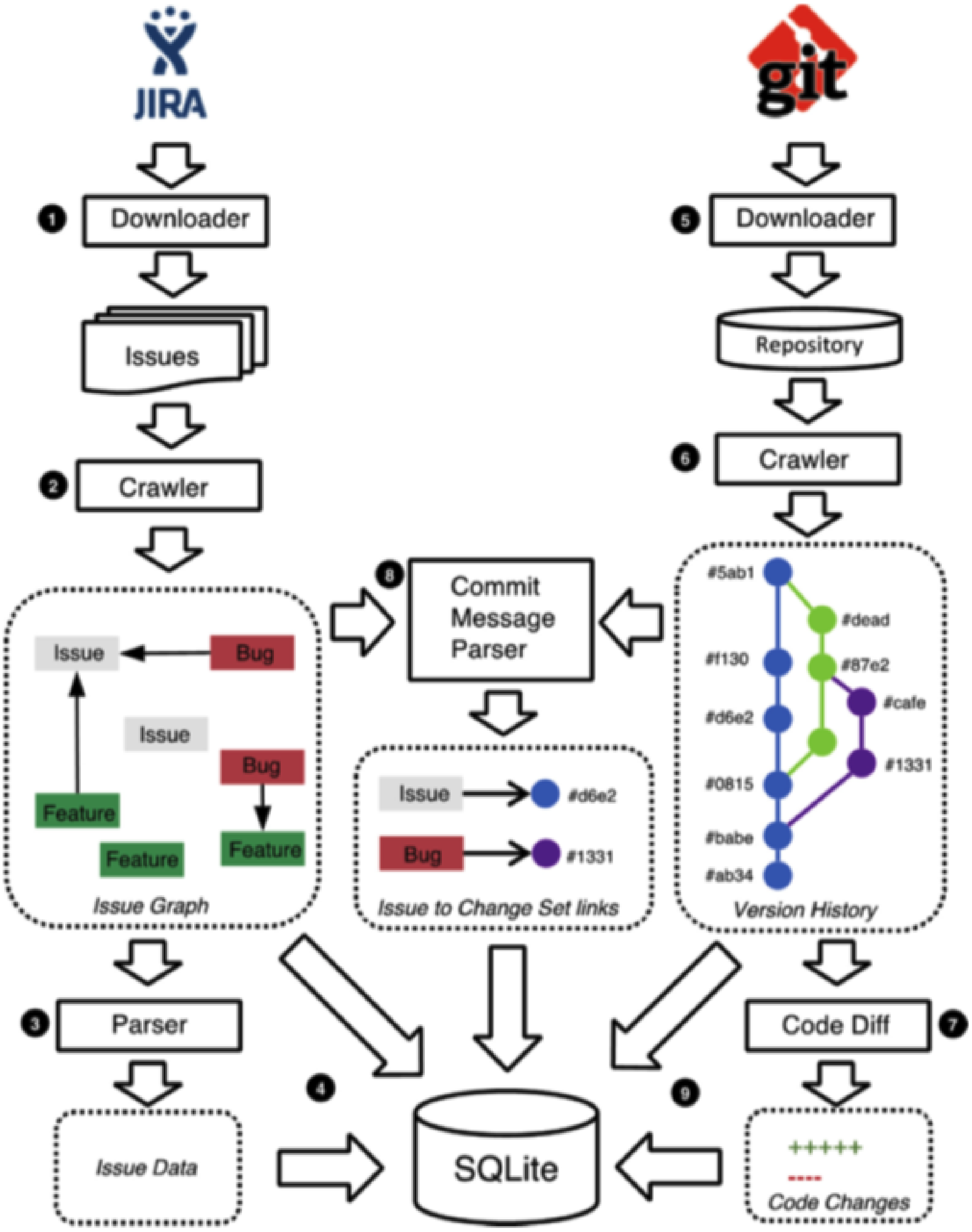


Figure 3.1. SEOSS33 data retrival

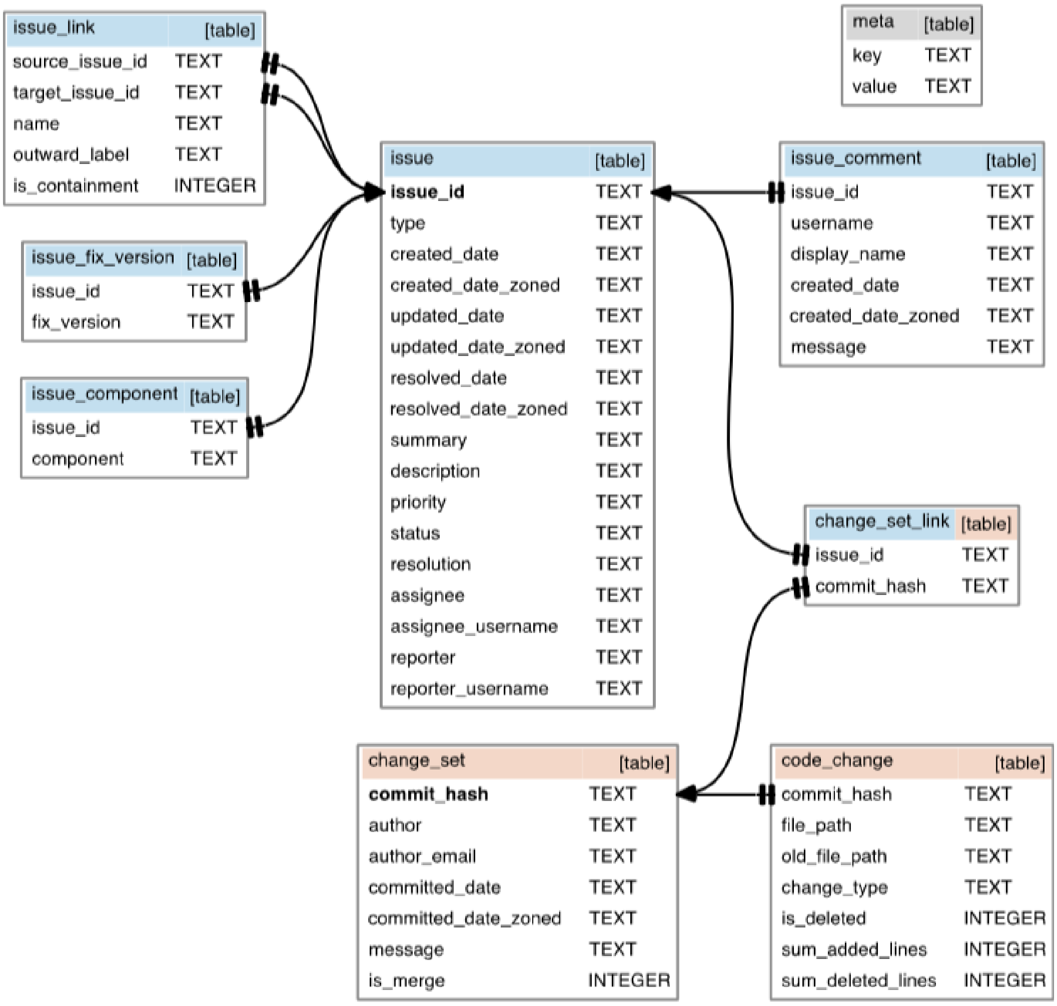
development for at least three years. Among these products we have selected five of them, because of size, as shown in table 3.1:

Table 3.1. Project data distribution

Project	**Month**	**Issue**
Hadoop	150	39086
Hbase	131	19247
Maven	183	18025
Cassandra	106	13965
Hive	113	18025

More selection charactheristics can be found in chapter 2.1 [1].
Is fundamental to understand the structure of this dataset, the majority of the forecasting operation tests are conducted using the data stored by this research. Each project is stored in a SQLITE file, a SQL offline database, the structure is based on the entity of the *issue*, identified by an *issue_id*, the other tables are used to link additional information, like the number of commit, the version referred, comments and others features. The figure 3.2 show the database schema.

Each table contains different types of data that will be used to generate the data used to forecast.



issue_link	[table]
source_issue_id	TEXT
target_issue_id	TEXT
name	TEXT
outward_label	TEXT
is_containment	INTEGER

meta	[table]
key	TEXT
value	TEXT

issue	[table]
issue_id	TEXT
type	TEXT
created_date	TEXT
created_date_zoned	TEXT
updated_date	TEXT
updated_date_zoned	TEXT
resolved_date	TEXT
resolved_date_zoned	TEXT
summary	TEXT
description	TEXT
priority	TEXT
status	TEXT
resolution	TEXT
assignee	TEXT
assignee_username	TEXT
reporter	TEXT
reporter_username	TEXT

issue_comment	[table]
issue_id	TEXT
username	TEXT
display_name	TEXT
created_date	TEXT
created_date_zoned	TEXT
message	TEXT

issue_fix_version	[table]
issue_id	TEXT
fix_version	TEXT

issue_component	[table]
issue_id	TEXT
component	TEXT

change_set_link	[table]
issue_id	TEXT
commit_hash	TEXT

change_set	[table]
commit_hash	TEXT
author	TEXT
author_email	TEXT
committed_date	TEXT
committed_date_zoned	TEXT
message	TEXT
is_merge	INTEGER

code_change	[table]
commit_hash	TEXT
file_path	TEXT
old_file_path	TEXT
change_type	TEXT
is_deleted	INTEGER
sum_added_lines	INTEGER
sum_deleted_lines	INTEGER

Figure 3.2. SEOSS33 data model

Chapter 4

Machine Learning

4.1 Introduction

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead[2]. The word ML is almost in the public domain now, in the last decades the usage of these kind of algorithms has drammatically raisen although most of it had already been developed for years. The main reason is the increase in the computational capacity of the systems.
There are two types of systems:

- Knowledge-based: Acquisition and modeling of common-sense knowledge and expert knowledge (from rules to facts)
- Learning: Extraction of knowledge and rules from example/experience (from facts to rules)

There a lot of different models available, the following chapters will focus on the models used in this project.

4.2 Supervised Learning

Random Forest Random Forest (RF) is a supervised learning algorithm which uses ensemble learning method for classification and regression. The forest is made by lot of different decision tree, its basic unit. The structure of the decision tree is simple, each branch define a direction to follow based on the values of different features, the end of a branch, the leaf, instead is the final predicted value. When all the trees are trained the model can be used, all the features value are evaluated by all the trees, than using some aggregation technique the final predicted value is computed. Using lot of different decision tree reduce the habit of overfitting. The behaviour is similar even for classification. Figure 4.1 shows a simplified schema of the model.

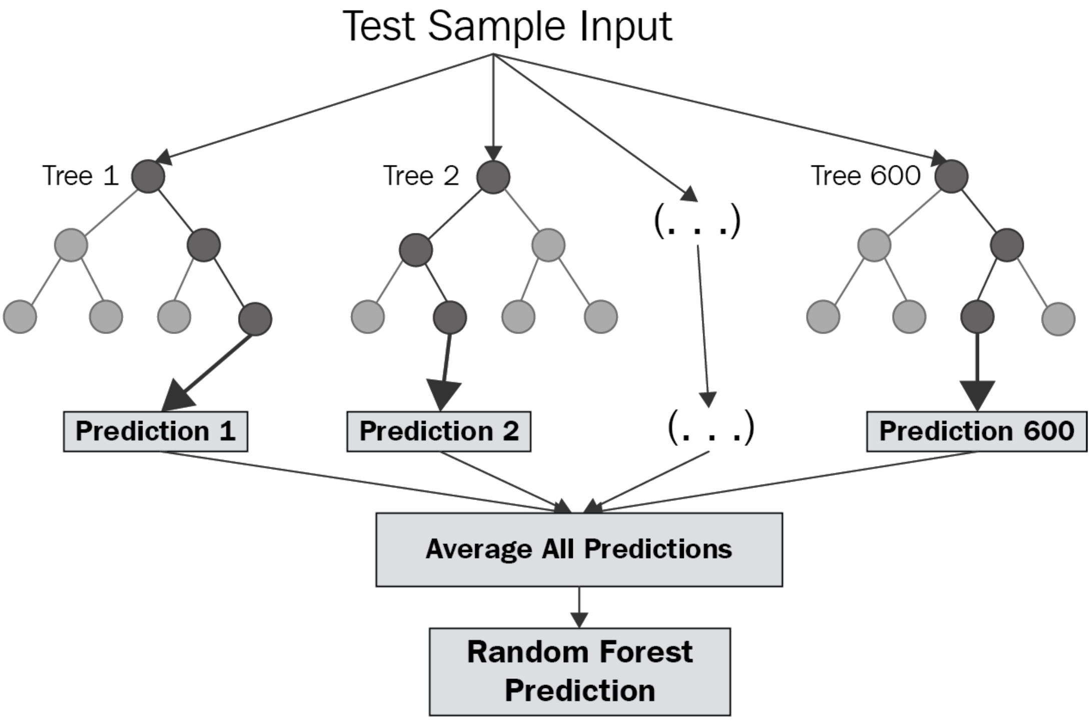


Figure 4.1. Random Forest simplified scheme [3]

The main advantages are that is efficent on large databases, can handle a lot of input variables without variable deletion, it generate an unbiased estimate of the generalization error for the build process and it can handle missing data. One of the drawback of this technique is the habit to overfit in particular condition, depending on datasets.

Gradient Boosting

4.3 Neural Networks

Neural Networks (NN) are algorithm, for pattern recognition, inspired by the structure of the human brain, with elaboration units (neurons) and connection network (synapses), exposed to enough of the right data, this kind of algorithms is able to establish correlations between present and future events. Figure 4.2 shows a simplyfied versiona of a neural network.
NN can be used to solve different kind of problems, classification, clustering and

regression. Our problem will be solved using the regressive type.
Regression analysis can be used to forecast one or more label based to other features. The structure of these networks can be really complicated and a whole thesis can be made upon this topic that, for this reason, will not more discussed.

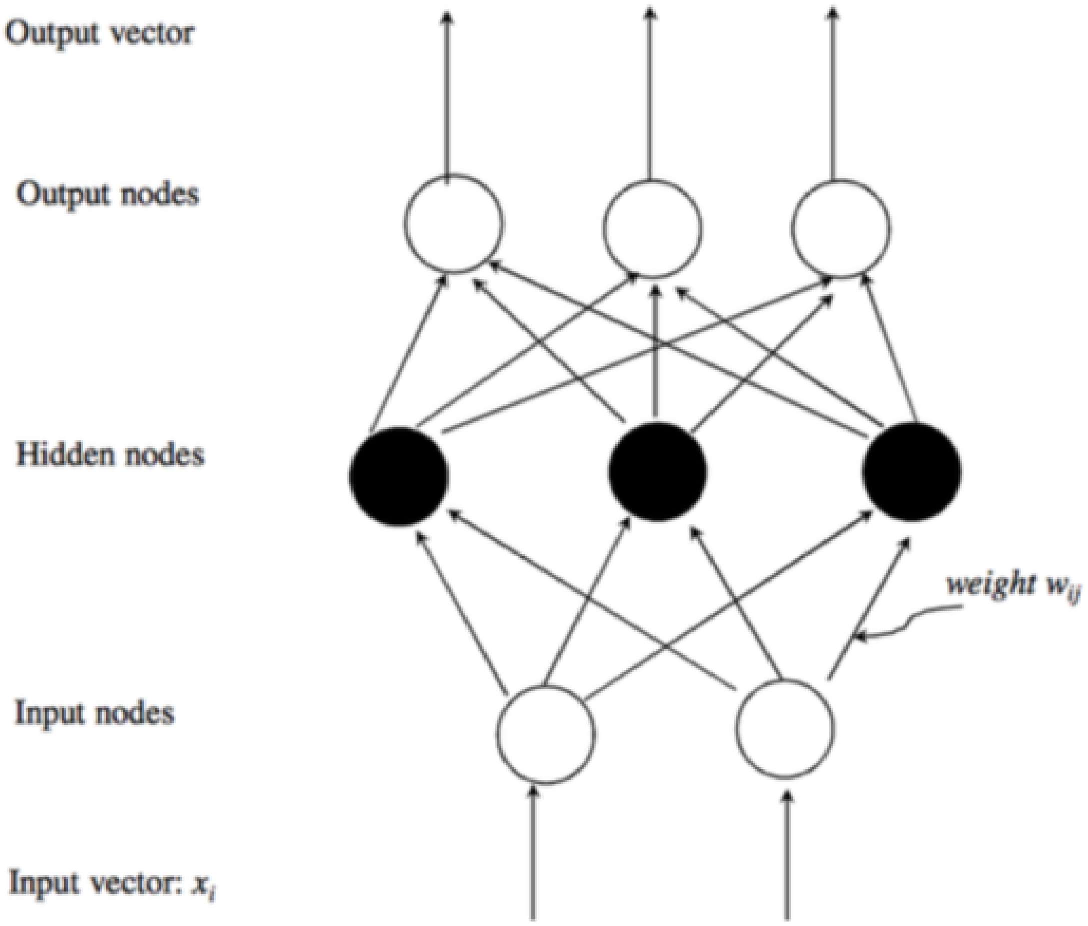


Figure 4.2. Structure of a neural network

Recurrent Neural Networks Recurrent Neural Networks (RNN) is a class of NN that keep connections between nodes and a temporal sequence. The main difference, respect NN, is that has feedback connections, this memory allow to keep track of temporal dynamic behaviour, they can process single data point or entire sequence of data, like video or speech.

Long-Short Term Memory One of the most used type of RNN is the Long-Short Term Memory (LSTM), were developed to solve the problems of exploding and vanishing of gradient typical of normal RNN.

Chapter 5

Forecasting

5.1 Introduction

Forecasting is the process of making predictions of future based on past and present data by trends analysis. Forecasting is one of the most desired machine learning functionality, it could be used to improve each kind of process, from finacials to production ones. Of course this task is not easy to achieve, a lot of resources and studies are needed to accomplish it. The software development is identical to a product development process, starts from the ideation and ends with the production itself. The goal is to predict the defectiveness in order to efficently allocate the development effort.

5.2 Features

The main advantage, in data analysis, of machines is that they can compute a lot of different data and finding a lot of patterns and correlation that human can't find. Combine the human attitude of logical correlations and machines capacity of number analysis can drive to a powerful combination that can drastically improve the forecasting ability. Each artificial intelligence algorithms require a correct and properly studied data in order to perform a valuable prediction, one of the basic step is the data preparation, providing correct and organized data is fundamental to correctly fit the network over the problem.

5.3 Models detail

5.4 One-Shot Prediction

5.5 Recurrent forecasting

5.6 Results

Chapter 6

Conclusion

Bibliography

[1] M. Rath, P. Mäder, “The SEOSS 33 Dataset — Requirements, Bug Reports, Code History, and Trace Links for Entire Projects” in *Data in Brief*, v. 25, p. 104005, 05 2019. [Online]: https://doi.org/10.7910/DVN/PDDZ4Q

[2] [Online]: https://en.wikipedia.org/wiki/Machine_learning

[3] [Online]: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f