PLELog

I. Introduction

PLELog is a novel approach for log-based anomaly detection via probabilistic label estimation. It is designed to effectively detect anomalies in unlabeled logs and meanwhile avoid the manual labeling effort for training data generation. We use semantic information within log events as fixed-length vectors and apply HDBSCAN to automatically clustering log sequences. After that, we also propose a Probabilistic Label Estimation approach to reduce the noises introduced by error labeling and put "labeled" instances into attention-based GRU network for training. We conducted an empirical study to evaluate the effectiveness of PLELog on two open-source log data (i.e., HDFS and BGL). The results demonstrate the effectiveness of PLELog. In particular, PLELog has been applied to two real-world systems from a university and a large corporation, further demonstrating its practicability.

This document is an introduction of how to install, test and reuse PLELog, a semi-supervised log-based anomaly detection approach, proposed in ICSE'21 accpeted paper "Semi-supervised Log-based Anomaly Detection via Probabilistic Label Estimation".

In this artifact submission, we are applying for a **Reusable Badge**.

We have submitted our research artifact including code, sample of experiment data and documentation. In this abstract document, we briefly introduce some key point of how to run PLELog, re-produce our experiment results and reuse PLELog on new log data.

This document is organized as follow: In sectionII, we introduce how to download PLELog and some key requirements of running PLELog. Later in Section III, we introduce how to use given script to verify the installation and re-produce our experiments in our paper. Finally in SectionIV, we introduce how to reuse PLELog on other log data for log-based anomaly detection.

II. Install

PLELog is a pure python project and currently archived in GitHub[1].

After cloning the repository, it is ready to run, but needs some environment preparation. The main packages including python 3, pytorch, hdbscan, overrides, scikit-learn. We have prepared a **requirements.txt** file for quick installation. One can easily install all required packages by simply runing:

pip install -r requirements.txt

[1]https://github.com/unksubmitter/PLELog

After the pip command finished, the environment should be ready to use.

III. Re-produce

A. Quick Start

To verify the installation and re-produce our experiment results, we prepared a trained model on BGL dataset along with all required data as a quick start sample for PLELog.

You can simply run the **test.py** script under the correct environment and, if all goes right, this script will start testing anomalies in BGL log data according to our experimental settings.

python test.py

When the script finishes, all results and related log are saved in the folder **log/test.log**, you can find the result from the end of the file.

IV. Re-Use

We have prepared an approach for anyone who are interested in PLELog and want to see if it is effective on their own data.

Please be noted that due to the indeterminacy of log format, to re-use PLELog, you must prepare a proper **Data Loading Function** as it will be able to read raw data according to your log format. For detailed instructions on how to implement PLELog on new log data is shown in Section Reproducibility in the README file.

V. CONCLUSION

In this research paper, we introduced a novel approach PLELog for log-based anomaly detection, and make it publicly available for future research. Currently, PLELog is still on its early version and we are planning on improving it continuously.

If you have any concerns or advice, please let us know. We are happy to be part of open-science and make some real influential research that actually helps resolving a challenge in software engineering practice.