IN4334 - MSR- Paper changelog

Eric Camellini, Kaj Dreef

January 9, 2016

1 Version 2 changes

1.1 Key improvements

The key improvements from version one to version two are:

- the paper was restructured following the received feedback;
- all the sections were rephrased accordingly to the received suggestions (apart for the related work);
- the paper now contains examples to make easier to understand the concepts of implicated files and the creation of the datasets;
- the paper now contains information about the projects that we use as study subjects and the size of the respective extracted datasets, together motivation for their selection;
- we provide the first results for every project;

1.2 Version 1 feedback

The feedback that we received on the first version was taken into account in the following way:

- we rephrased the introduction where requested;
- we rephrased the problem section, expanding the part that talks about previous solutions and related limitations;
- we merged the background and methodology sections;
- we then split the methodology section in two sections extracting a proposed solution section from it;
- we rephrased the research questions and we moved the researh questions at the beginning of the (new) methodology section;
- we are still considering only Random Forest, no logistic regression yet;

- we added more motivations in the study subjects section, and we moved it at the beginning of the methodology;
- we removed the part about the correlation and we now only speak about classification;

1.3 Intermediate analysis presentation feedback

The feedback that we received on the intermediate analysis presentation was taken into account in the following way:

- the authorship metrics are now included;
- we added the #ofPrevious bugs classic metric (that in our case counts the # of previous implications);
- we performed statistical significance tests on the improvement over classic;
- at the moment, for how we rephrased the research questions, we don't plan to change the threshold using features of the projects: we compute the dataset for many thresholds and check the difference.