Weighted Data Set Reduction for Bug Triaging

Miaomiao Wei, Shikai Guo and Rong Chen

Dalian Maritime University, Dalian, China
{weimiaomiao, shikai.guo, rchen}@dlmu.edu.cn

***Abstract*—Despite the great potential to save the labor cost of developers, automated bug triaging as a text classification problem has not been thoroughly investigated on long descriptions, which are informative but often noisy. In this paper an effective bug triage technique is proposed to build a high quality set of bug data by removing the noisy and non-informative bug reports while assigning new bugs to an appropriate developer. The proposed technique – weighted data set reduction – is built upon three feature selection algorithms and four instances selection algorithms with intention to recommend the bug and to automatically assign it more accurately even with noisy bug descriptions. Several experiments are conducted and the experimental results show that the reduced training sets by the proposed approach can achieve better accuracy in several cases, about 2-3% on average better than the original ones.**

Keywords—Bug Triaging; Bug Reports; Machine Learning;

I. INTRODUCTION

With the huge information about bugs reported [1] by today's bug tracking systems (e.g. , Buzilla, JIRA, mantis), there is an increasing need to introduce some form of automation within the bug triaging process, so that no time is wasted on the assignment of new issues to appropriate developers. In this paper, we propose a weighted data reduction technique that combines feature selection and instance selection algorithms to yield a small size and high-quality training set that can enhance mainstream classifiers.

In doing so, three feature selection algorithms (CHI, Information Gain (IG), and One Rule (OneR)) can distinguish important attributes with meaningless ones, while four instance selection algorithms ICF, Condensed Nearest Neighbor (CNN), Minimal Consistent Set (MCS) and Edited Nearest Neighbor (ENN) can choose more representative examples.

II. MODEL

Figure 1 shows our bug triaging model based on data set reduction and text classification. The present approach is conducted in four steps: (1) Part of bug reports with the label are selected from the original data. (2) Data preparation. The selected bug reports are transformed into standard data set. (3) Data set reduction. The data set is reduced with feature selection and instances selection algorithms. (4) Classification. When a new bug report is submitted, a size of K recommendation list is generated by a concrete classifier.

Before data set reduction, the data are prepared as follows: first, select the bug not empty. Second, select the fixed bugs and duplicate bugs. Third, label the bugs and delete the console information. Fourth, remove the inactive developers. Fifth, separate words and remove stop words. At last, generate a feature matrix as the standard data.

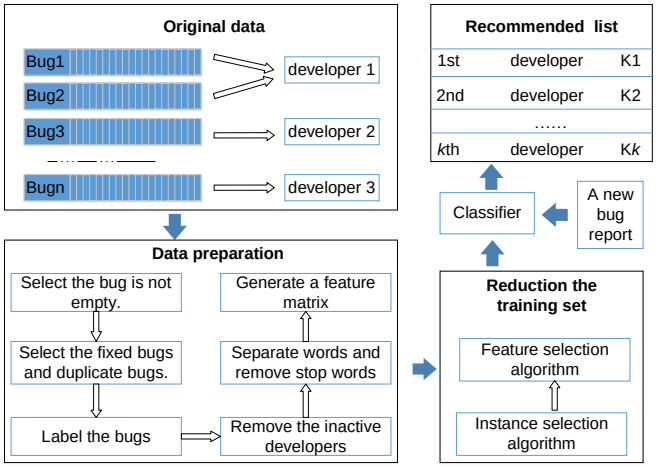

Figure 1. The text categorization approach for bug triaging.

Algorithm 1. Weighted Reduction training set algorithm.

Input: training set X,
Output: X^*

1. for each bug report in training set
2. for each word in short description
3. word frequency* η
4. end for
5. end for
6. generate a weighted feature matrix
7. while not approach the scale of reduction
8. apply *FS →IS or IS →FS to* weighted feature matrix
9. return X^*

III. EXPERIMENT AND CONCLUSION

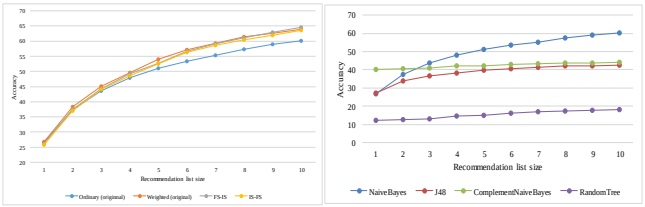

Figure 2 the results of dataset reduction.

Figure 2 shows the results of our approach. NB classifier has the best triaging effect in our study compared with the other three bug classification algorithms. Therefore, FS-IS is a good choice for the training set reduction.

REFERENCES

[1] Zhang J, Wang X Y, Hao D, et al, "A survey on bug-report analysis," Science China Information Sciences, 2015, 58(2):1-24.

DOI:10.18293/SEKE2018-053