An Empirical Study of Bugs in Machine Learning Systems

Ferdian Thung, Shaowei Wang, David Lo and Lingxiao Jiang

Abstract: Many machine learning systems that include various data mining, information retrieval, and natural language processing code and libraries have been used in real world applications. Search engines, internet advertising systems, product recommendation systems are sample users of these algorithm intensive code and libraries. Machine learning code and toolkits have also been used in many recent studies on software mining and analytics that aim to automate various software engineering tasks. With the increasing number of important applications of machine learning systems, the reliability of such systems is also becoming increasingly important. A necessary step for ensuring reliability of such systems is to understand the features and characteristics of bugs occurred in the systems. A number of studies have investigated bugs and fixes in various software systems, but none focuses on machine learning systems. Machine learning systems are unique due to their algorithm-intensive nature and applications to potentially large-scale data, and thus deserve a special consideration.

In this study, we fill the research gap by performing an empirical study on the bugs appeared in machine learning systems. We analyze three systems, namely Apache Mahout, Lucene, and OpenNLP, which are data mining, information retrieval, and natural language processing tools respectively. We look into their bug databases and code repositories, analyze existing bugs and corresponding fixes, and label the bugs into various categories. Our study finds that 22.6% of the bugs belong to algorithm/method category, 15.6% of the bugs belong to the non-functional category, and 13% of the bugs belong to the assignment/initialization category. We also report the relationship between the categories of bugs and their severity, the time and effort needed to fix the bugs, and their impact. We highlight several categories of bugs that deserve attention in future research.