On the Classification of Software Change Messages using Multi-label Active Learning

Sirine Gharbi  
LARODEC - Institut Supérieur de Gestion de Tunis  
41, Avenue de la liberté, 2000, Le Bardo, Tunisia  
gharbi.cyrine@outlook.com

Mohamed Wiem Mkaouer  
Rochester Institute of Technology  
Rochester, NY, USA  
mwmvse@rit.edu

Ilyes Jenhani  
Prince Mohammad bin Fahd University  
Khobar, Saudi Arabia  
ijenhani@pmu.edu.sa

Montassar Ben Messaoud  
LARODEC - Institut Supérieur de Gestion de Tunis  
41, Avenue de la liberté, 2000, Le Bardo, Tunisia  
montassar.benmassaoud@isgs.u-sousse.tn

ABSTRACT
In this paper, we present a multi-label active learning-based approach to handle the problem of classification of commit messages. The approach will help developers track software changes, e.g., adding or updating existing features, fixing user-reported errors, improving software performance, etc. We first constructed an unlabeled dataset of commit messages where each commit message is represented as a vector of feature values. The set of adopted features were automatically generated from the original commit messages using Term Frequency-Inverse Document Frequency (TFIDF) Technique. Because many commit messages can be assigned more than one commit class at the same time and in order to reduce the effort needed to assign labels to each instance in a large set of commit messages, we adopted an Active Learning multi-label approach. Experimentations have shown that we could train an accurate multi-label classifier model, in our case, a binary relevance with logistic regression as a base classifier, by actively querying an oracle for labels during the training process and with a reasonable number of labeled instances.