Syntactic and Semantic Feature Extraction and Preprocessing to Reduce Noise in Bug Classification

Ruchi Agrawal and G. Ram Mohan Reddy

National Institute of Technology, Karnataka, India
{agrawalruchi01,profgrmreddy}@gmail.com

Abstract. In software industry a lot of effort is spent in analyzing the bug report to classify the bugs. This Classification helps in assigning the bugs to the specific team for Bug Fixing according to the nature of the bug. In this paper, we have proposed a data mining technique applying syntactic and semantic Feature Extraction to assist developers in bug Classification. Extracted features are organized into different feature groups then a specific preprocessing technique is applied to each feature group. The applied methods have reduced the noise in the bug data compared to traditional approach of word frequency for text categorization. We have analyzed our approach on a collection of bug reports collected from a networking based organization (CISCO).The experiments are performed using Naive Bayes Multinomial Model and Support Vector Machine on features obtained after preprocessing.

Keywords: Bug Fixing, Classification, Feature Extraction, Naive Bayes, Support Vector Machine.

1 Introduction

Large organizations like CISCO require a bug classifier system. Since these organization are having many products and different maintenance team to handle different types of bugs. Assigning a bug to a particular team, so that it can be resolved quickly is a challenging task. This system helps to classify the bugs according to different maintenance teams of the organization. Thus, it aims to reduce the overall time to fix the bug.

Like most of the big organizations, CISCO is also having its own bug tracking system which contains bug information in form of various attachments. Attachments refer to the links for accessing the data regarding bug such as description, crash log info, and stack trace decode and other information. It also has provision to add comments and information after static analysis of the bug report as a separate attachment. This system of posting comments is also similar to most open source bug tracking system like bugzilla [1].

The attachments added manually such as description and static analysis are in natural language format (semantic information) whereas the crash log file collected from the crashed system contain information in programming language

K.R. Venugopal and L.M. Patnaik (Eds.): ICIP 2012, CCIS 292, pp. 329–339, 2012.

format (syntactic information). Instead of using traditional approach of word frequency for text categorization, information from the attachments can be mined to find out some specific pattern for Feature Extraction and classification. This paper aims at reducing the noise in the data so that bugs can be classified correctly and quickly.

In this paper, we analyze the network bugs and depending on the static analysis of the bug report, the Feature Extraction is performed .The features are grouped into different feature groups and different preprocessing technique is applied to the extracted features to reduce the level of noise in the data. Any classification approach can be applied on the extracted features; we had analyzed our approach using Bayesian probability approach and Support Vector Machine.

2 Related Work

Davor Cubranic et.al. [9] have proposed an approach for automatic bug triage using text categorization. They proposed a prototype for bug assignment to developer using supervised Bayesian learning. Their prototype used the word frequency as input to the classifier. In our approach instead of considering word frequency we had taken bug semantics into consideration. Our approach helps to reduce noise in the extracted features.

Nicholas et.al. [8] have proposed a system that automatically classifies duplicate bug reports as they arrive to save developer time. Their system used surface features, textual semantics, and graph clustering to predict duplicate status. They had considered only textual features that are title and description. In our approach ,syntactic features along with the textual features are used to increase the accuracy of classification.

Deqing Wang et.al [7] have implemented a tool Rebug-Detector, to detect related bugs using bug information and code features. The extracted features related to bugs and used relationship between different methods that is overloaded or overridden methods. In our approach we had used the sequence of the function call present in the stack at the time crash happened. Since the stack image is present for all the bugs irrespective of the product or organization, our approach can be applied on any bug database.

Karl-Michael Schneider in the paper [5] used Naive Bayes Method for Spam Classification. Kian Ming Adam Chai, Hwee Tou Ng and Hai Leong Chieus in their paper [6], explores the use of Bayesian probability approach for text classification. They showed through experiments that Bayesian is good approach for text classification. The words can be considered as unigram features obtained irrespective of the type of bugs.

3 Feature Extraction and Preprocessing

3.1 Overview of the Bug Site

In bug site, bug reports are organized in the form of different attachments and attachments are grouped into General, Commit, Build, Test, Fix Entries category.