Early Identification of Novice Programmers’ Challenges in Coding Using Machine Learning Techniques

Alireza Ahadi
University of Technology, Sydney
Sydney, NSW, 2008, Australia
Alireza.Ahadi@uts.edu.au

ABSTRACT
It is well known that many first year undergraduate university students struggle with learning to program. Educational Data Mining (EDM) applies machine learning and statistics to information generated from educational settings. In this PhD project, EDM is used to study first semester novice programmers, using data collected from students as they work on computers to complete their normal weekly laboratory exercises. Analysis of the generated snapshots has shown the potential for early identification of students who later struggle in the course. The aim of this study is to propose a method for early identification of “at risk” students while providing suggestions on how they can improve their coding style. This PhD project is within its final year.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education – computer science education.

Keywords
Snapshot Analysis; Novice Programmers; Machine Learning

1. CONTEXT AND MOTIVATION
Every year, tens of thousands of students fail introductory programming courses world-wide. As a consequence, studies are retaken or postponed, careers are reconsidered, and substantial capital is invested into student counseling and support. Worldwide, on average one third of students fail their introductory programming course. Even when looking at statistics describing pass rates after teaching interventions, as many as one quarter of programming course world. Every year, tens of thousands of students fail introductory programming courses. Watson et al. noted that a simple measure, the average amount of time that a student spends on programming their programming err.ors, if there is any is an environmental independent common attribute of the data collected from novices in different programming languages which could automatically classify students according to their ability of coding?

The research goals are as follows:

Data Collection: data used in this study is collected from source code snapshots generated by novices enrolled in an introductory programming courses at a) University of Helsinki, Finland. b) University of Technology, Sydney, and c) snapshots generated by novices enrolled in an database fundamental course at University of Technology, Sydney. The collected programming snapshots represent line-edit level [4] snapshots of the students’ main source code while working on a programming task, while the database SQL SELECT statements represent different attempts of novices in writing SQL code.

Replication: analyzing the collected data in an unsupervised machine learning framework to assess the state of the art techniques described above. The collected source code snapshots
will be used to identify possible clusters of novices according to different features extracted from their source codes.

**Proposition and Comparison:** analyzing the collected data in order to compare the performance of new methods proposed as a part of this PhD with the state of the art methods. Based on the output of different machine learning tools, a set of context-independent features extracted from the data as well as a systematic approach for analyzing them will be proposed.

4. **PROGRAM CONTEXT**

Accomplishing the final output of this research project involves fulfilling the objectives of three main phases. The first phase is dedicated to analysis of snapshot data in different contexts [5]–[9] to understand the data and find possible links to what previously has been reported [10]–[16]. The second phase is devoted to designing machine learning derived techniques for snapshot data analysis [6]. The third and last phase of the project is dedicated to performing the comparative analysis of the proposed techniques of analyzing snapshot data with the state of the art techniques in a large scale on different datasets.

5. **DISSERTATION STATUS**

This PhD project is within a year of completion. The data collection has been completed successfully and some preliminary results have been generated based on an original hypothesis. Replication and comparison stages are almost completed. Based on findings so far, there is a significant and strong negative correlation between the number of steps a novice takes to complete a coding task successfully and her overall performance in the programming subject.

6. **REFERENCES**


