Computers in Human Behavior 73 (2017) 247–256

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses

Evandro B. Costa [a, *], Baldoino Fonseca [a], Marcelo Almeida Santana [a], Fabrísia Ferreira de Araújo [b, c], Joilson Rego [d]

[a] Federal University of Alagoas (UFAL), Brazil
[b] Federal Institute of Alagoas (IFAL), Brazil
[c] Federal University of Campina Grande, Brazil
[d] Federal University of Rio Grande do Norte (UFRN), Brazil

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form
16 January 2017
Accepted 26 January 2017
Available online 4 February 2017

Keywords:
Artificial intelligence in education
Automatic instructional planner
Automatic prediction
Educational data mining
Interactive learning environment
Learner modeling

ABSTRACT

The data about high students' failure rates in introductory programming courses have been alarming many educators, raising a number of important questions regarding prediction aspects. In this paper, we present a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fail in introductory programming courses. Although several works have analyzed these techniques to identify students' academic failures, our study differs from existing ones as follows: (i) we investigate the effectiveness of such techniques to identify students likely to fail at early enough stage for action to be taken to reduce the failure rate; (ii) we analyse the impact of data preprocessing and algorithms fine-tuning tasks, on the effectiveness of the mentioned techniques. In our study we evaluated the effectiveness of four prediction techniques on two different and independent data sources on introductory programming courses available from a Brazilian Public University: one comes from distance education and the other from on-campus. The results showed that the techniques analyzed in our study are able to early identify students likely to fail, the effectiveness of some of these techniques is improved after applying the data preprocessing and/or algorithms fine-tuning, and the support vector machine technique outperforms the other ones in a statistically significant way.

1. Introduction

The alarming indexes of students' academic failures, along the years, in universities' introductory programming courses (Bennedsen & Caspersen, 2007; Watson & Li, 2014) have been concerning educators. Studies (Hanks et al., 2004; Iepsen et al., 2013; Tan, Ting, & Ling, 2009) show that students face many difficulties during their programming activities in such a way that many of them end up failing or quitting the course at some initial stage.

In the above context, one relevant problem is on the ability to accurately predict the students likely to fail in introductory programming courses at early enough stage for possibiliting pedagogical interventions to be taken to avoid students' failures. In order to deal with this problem, some works (Arora, Singhal, & Bansal, 2014; Bayer et al., 2012; Manhães et al., 2014; Marquez-Vera, Morales, & Soto, 2013; Martinho, Nunes, & Minussi, 2013; Sim et al., 2006; Watson, Li, & Godwin, 2013) have proposed and analyzed the use of Educational Data Mining (EDM) techniques to predict students' academic failures. However, in general, these works are not concerned with two important questions: (i) how effective are the EDM techniques to early identify students likely to fail?; and (ii) Do the data preprocessing (Hu, 2003; Crone et al., 2006; Zaki & Jr.W. M., 2014) and algorithms fine-tuning (Gunawan et al., 2011; Hutter, Hoos, Leyton-Brown, & Stützle, 2009) impact the effectiveness of EDM techniques?

In order to answer the above mentioned questions, we present a comparative study on the effectiveness of EDM techniques to early predict students likely to fail in introductory programming courses. Given the amount of existing EDM techniques available (Caruana &

* Corresponding author.
E-mail addresses: evandro@ic.ufal.br (E.B. Costa), baldoino@ic.ufal.br (B. Fonseca), marceloalmeidasantana@gmail.com (M.A. Santana), fabrisia.araujo@gmail.com (F.F. de Araújo), jotarego@gmail.com (J. Rego).

http://dx.doi.org/10.1016/j.chb.2017.01.047
0747-5632/

Niculescu-Mizil, 2006), we used the following classifiers: Neural Networks (Nürnberger et al., 2002; Rumelhart, Hinton, & Williams, 1988), Decision Tree (Breiman et al., 1984; Salzberg, 1994), Support Vector Machine (SVM) (Cortes & Vapnik, 1995; Vapnik, 1995) and Naive Bayes (Domingos & Pazzani, 1997). These techniques have been widely investigated by existing EDM works (Wu et al., 2008) and they have presented interesting results.

In our study we used the f-measure (Han et al., 2011) to evaluate the effectiveness of the selected techniques on two different and independent data sources concerning two introductory programming courses available from a Brazilian Public University: one comes from distance education and the other from on-campus. The experiment was performed by considering the preprocessing of these data sources and the fine-tuning of the analyzed techniques.

The results showed that the techniques analyzed in our study are able to early identify students likely to fail, and demonstrated that the data preprocessing and algorithms fine-tuning tasks influence the effectiveness of these techniques. The SVM technique outperformed the other ones by predicting with 92% and 83% of effectiveness the failures of students that have performed at least 50% of the courses by distance education or on-campus, respectively.

This paper is organized as follows. Section 2 we present the method applied in our experiment. In Section 3 we present the results and discussions of the experiment. In Section 4 we discuss some similar work. Conclusions and future work are presented in Section 5.

2. Method

The general goal of this study is to compare the effectiveness of existing EDM techniques for early identification of students likely to fail with high precision. This section is organized as follows. Section 2.1 poses four research questions that drive our assessment. Section 2.2 presents the data sources and the EDM techniques we have analyzed in our experiment. Sections 2.3 and 2.4 indicate the tools and metrics, respectively, we have used when conducting the experiment, and, finally, Section 2.5 presents some details about the steps and configurations used to perform the experiment.

2.1. Planning

Our comparative study is guided by the following research questions:

Question 1. How effective are the EDM techniques to early identify students likely to fail?

Our aim with the Question 1 is to evaluate the effectiveness of the EDM techniques that have been used by existing approaches to early identify students likely to fail. To answer Question 1, we performed these techniques on two different data sources and then we used the F-measure to evaluate the effectiveness of such techniques.

Question 2. Is the data preprocessing able to increase the effectiveness of the EDM techniques?

The **Question 2** aims to analyse if the effectiveness of EDM techniques increases after performing the data preprocessing. In order to answer **Question 2**, we performed a preprocessing of the two data sources used in this experiment, then we applied the EDM techniques on these data sources. Subsequently, we evaluated the effectiveness of these techniques and we compared such results with the effectiveness obtained by performing the same techniques on the data without the preprocessing.

Question 3. Is the fine-tuning of algorithms able to further increase the effectiveness of the EDM techniques?

The **Question 3** aims to analyse if the effectiveness of EDM techniques further increases after performing the fine-tuning of their parameters. In order to answer **Question 3**, we performed a fine-tuning of the EDM techniques, then we performed the fine-tuned techniques on the preprocessed data source, as well as we evaluated the effectiveness of techniques and we compared their effective-ness with the results obtained by performing the EDM techniques without the fine-tuning.

Question 4. After performing the data preprocessing and fine-tuning of algorithms, which of the EDM techniques are more effective for early identification of students likely to fail?

The **Question 4** aims to find the most effective techniques for early identification of students likely to fail. In order to answer **Question 4**, we analyzed the effectiveness of the EDM techniques after performing the fine-tuning of their parameters and the preprocessing of the data sources.

2.2. Data sources and EDM techniques selection

In this experiment we have analyzed two data sources extracted from introductory programming courses performed either on-campus or distance education. In what follows a brief description of these two data sources:

(Distance Education) The first data source contains information about 262 undergraduate students that took the introductory programming course performed in a distance education modality in our university in 2013 during 10 weeks. In this course the students were weekly evaluated according to their activities plus two exams that were applied in the fifth and last week of the course. These activities and exams were applied through an online system used in our university.

This data source contains the following information about the students: age, gender, civil status, city, income, student registration, period, class, semester, campus, access frequency of the students in the system, participation in the discussions forum, amount of received and viewed files, use of the educational tools provided by the system as blog, glossary, quiz, wiki, message, year of enrolling in the course, status on discipline, and performance of the students in the weekly activities and exams.

(On-campus) The second data source contains information about161students that took the introductory programming course performed on-campus in our university in 2014, during 16 weeks. In this course the students were weekly evaluated according to their activities plus four exams that were applied in the fourth, eighth, twelfth and sixteenth week of the course.

The data source contains the following students information: age, gender, civil status, city, income, student registration, period, class, semester, campus, year of enrolling in the course, status on discipline, amount of exercise performed by the student, number of correct exercises, and performance of the students in the weekly activities and exams.

Our main goal is to evaluate the effectiveness of the EDM techniques to predict students likely to fail at early enough stage for supporting future pedagogical interventions to be taken to avoid