Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 614–621

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes

Shinichi Oeda[a*], Genki Hashimoto[b]

[a]*Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College*
[b]*Advanced Control and Information Engineering Course, National Institute of Technology, Kisarazu College, 11-1, Kiyomidaihigashi 2-chome Kisarazu City, Chiba, Japan*

Abstract

Educational data mining (EDM) involves the application of data mining, machine learning, and statistics to information generated from an educational setting. In most school education, one teacher teaches many students. A periodic examination is used as a method to confirm that students have acquired skills. However, it is difficult to grasp the status of the student from each lesson, since examinations cannot be carried out easily. On the other hand, in programming classes, the students' history of UNIX commands and source-code editing can be easily and automatically stored as log-data. Therefore, attempts have been made to estimate the student's performance from this log-data, although their estimation accuracy is not high. In this research, we aim to extract those students who cannot keep up with programming lessons, rather than estimating the student's performance from the log-data. Specifically, we propose a method for predicting dropouts using outlier detection to cluster data with unsupervised learning.

Peer-review under responsibility of KES International

Keywords: Educational data mining, Dropout prediction, Dynamic time warping, k-means++, UNIX command history

1. Introduction

Intelligent tutoring systems (ITSs) and learning management systems (LMSs) have been widely used in the field of education, and have allowed us to collect log-data from learners, and especially students. Educational data mining (EDM) aims at discovering useful information from massive amounts of electronic data collected by these educational systems. EDM is an emerging multi-disciplinary research area, with several methods and techniques developed to explore data originating from various educational information systems [1].

We focus on the actual school educational system. In most schools, one teacher teaches many students in a class at a time. It is thus difficult for teachers to provide detailed guidance for each respective student. Indeed, a teacher needs

* Corresponding author. Tel.: +81-438-30-4144 ; fax: +81-438-98-5717.
E-mail address: oeda@j.kisarazu.ac.jp

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.088

to know the students' level of skill in order to provide high-quality education. Because of this, periodic examinations are used as a method to confirm whether students have acquired the necessary skills. However, it is difficult to grasp the status of the student from each lesson, since examinations are time consuming and burdensome to grade. On the other hand, it is easy to automatically record behavior in programming classes, since the history of each student's UNIX commands and source-code editing can stored as log-data.

There are several studies available regarding the use of this log-data. For instance, there is research into predicting student's ability from log-data[2]. However, the prediction accuracy of such methods is not high, and they do not consider whether the verification of the grading evaluation function is insufficient. Moreover, some research suggests that it is difficult to evaluate students' acquired skills from log-data.

Therefore, in this research, we aim to extract those students who cannot keep up with programming lessons, rather than estimating the student's performance from log-data. Specifically, we propose a method for predicting dropouts using outlier detection to cluster data with unsupervised learning.

2. Problem Setting for Dropout Prediction

Monitoring and supporting students is considered very important at many educational institutions. If a teacher can detect a weak student at an early stage, he or she can take measures to ensure that the student will not drop out of the class. Therefore, it is necessary to predict those students at risk of dropping out during each lesson, so that the teacher can pay special attention to them. In programming classes, it is possible to confirm the behavior of students using log-data. We developed a logging system that can acquire the input history of UNIX commands and the editing history of source code. We obtained several datasets from students in our school. Our dataset from programming lessons consists of data from 39 students. When predicting potential dropouts, it is common to create evaluators with supervised learning[3,4,5]. However, such data is difficult to learn, insofar as the size of our dataset is small. Moreover, the features in the log-data depend on the content of each lesson, on whether there are multiple exercises or an abundance of explanations. Therefore, we apply outlier detection with unsupervised learning. We assume that students who belong to the outlier cluster can be compared with either excellent students or inferior ones. We performed k-means clustering according to the Euclidean distance, in order to compare the timing of active behavior and clustering using dynamic time warping. Thus, we can compare the flow of activity to the exclusion of time-series deviations.

Massive Open Online Courses (MOOCs) offer a new and increasingly popular model for delivering educational content online to any person who wants to take a course, with no limit on attendance. MOOCs are a recent and widely researched development in distance education, first introduced in 2006. Already by 2012, MOOCs emerged as widely popular mode of learning[6,7]. However, because the number of students in our target classes is fewer than 50, a new model must be developed. In Japan, there are several studies pertaining to the improvement of classes using log-data analysis. In most of these studies, the researchers developed original tools in order to obtain log-data based on the behavior of students, and analyzed this data by writing programs using the tools[11,12]. By contrast, we aim at analyzing log-data that can be acquired without affecting the coding environment of the students and without educational constraints.

3. Dynamic Time Warping

Dynamic time warping (DTW)[8] is an algorithm that compares all points in two time series with each other, and sets the combination of optimal points as a distance. With DTW, it is possible to compare time series of different lengths, corresponding to phenomena shifted in the time series direction, such as delays and enlargements. The algorithm for finding the distance of an M-dimensional time series vector $\boldsymbol{x}_M$ and an N-dimensional time series vector $\boldsymbol{x}_N$ is as follows.

1. $M \times N$ matrix $\boldsymbol{D}$ is given $\boldsymbol{D}_{1,1} = 0$, and other elements are given ∞.
2. Calculate $\boldsymbol{D}_{i,j}$ for all $i(= 2, \cdots, N)$, $j(= 2, \cdots, M)$, as follows,

$$\boldsymbol{D}_{i,j} = \sqrt{(x_{Ni}, x_{Mj})^2} + \min(\boldsymbol{D}_{i-1,j}, \boldsymbol{D}_{i,j-1}, \boldsymbol{D}_{i-1,j-1}) \tag{1}$$