Chapter 5
Sequence Clustering Techniques in Educational Data Mining

Qi Guo
Medical Council of Canada, Canada

Ying Cui
University of Alberta, Canada

Jacqueline P. Leighton
https://orcid.org/0000-0001-8958-2470
University of Alberta, Canada

Man-Wai Chu
University of Alberta, Canada

ABSTRACT

Digital technology has profound impacts on modern education. Digital technology not only greatly improves access to quality education, but it also can automatically save all the interactions between students and computers in log files. Clustering of log files can help researchers better understand students and improve the learning program. One challenge associated with log file clustering is that log files are sequential in nature, but traditional cluster analysis techniques are designed for cross-sectional data. To overcome this problem, several sequence clustering techniques are proposed recently. There are three major categories of sequence clustering techniques: Markov chain clustering, sequence distance clustering, and sequence feature clustering. The purpose of this chapter is to introduce these sequence clustering techniques and discuss their potential advantages and disadvantages.

DOI: 10.4018/978-1-7998-3476-2.ch005

INTRODUCTION

Recent advances in computer technology have offered opportunities for new ways of measuring complex performance and higher order skills using dynamic, interactive, and real-world-like tasks. Computer technology also makes it possible to record every problem solving action a student made in a log file. Analysis of log files can help researchers and instructors understand how students arrive at a conclusion. It also helps personalize the e-learning environment for each student, and provides specific automatic feedbacks to students. To analyze log files, student problem solving sequences often need to be clustered. However, due to log files' complexity, there are several challenges for applying traditional cluster analysis. Kerr (2015) summarized the literature and pointed out three major challenges for clustering log files. First, log files contain large quantity of information. Second, log files contain diverse actions, not all are necessarily related to the problem. Third, students' log files may have little overlap as different students may do completely different behaviors. In addition to the above three challenges, it is also important to note that log file information is sequential in nature. That is, doing action A and then doing action B may be different from doing action B first and then doing action A, even though the same actions are performed. The orders in which students perform certain actions may be crucial to understand their problem solving strategies.

Traditional cluster analysis methods designed for non-sequential problems cannot be directly applied to the clustering of log data. To overcome the problem, many sequence clustering algorithms were proposed. These algorithms can be grouped into three major approaches: Markov chain clustering approach (Hansen, Hansen, Hjuler, Alstrup, & Lioma, 2017), sequence distance clustering approach (Shen & Chi, 2017), and sequence feature clustering approach (Wang, Weng, Su, & Tseng, 2004). The goal of this paper is to introduce these approaches and demonstrate each approach with some recently proposed algorithms. In the background section, basic cluster analysis algorithms are introduced to provide a foundation for the main focus of the paper. Next, the main ideas of the three sequence clustering approaches are explained. To demonstrate each approach, several example algorithms are presented. In the solution and recommendation section, practical recommendations are given to guide when to use what approach. In the end, future research directions and significance of the topic are discussed.

BACKGROUND

Throughout the modern history of education, educators always attempt to make education more realistic, more performance-based, and more like one-on-one coaching (Mayrath, Clarke-Midura, & Robinson, 2011). Instead of letting students remember knowledge, educators want them to apply knowledge to solve real world problems. Instead of simply giving students a single test score, educators want to give students unique feedback, and recommend them to learn what they need to learn. However, in reality, educators simply do not have enough time and resources to achieve these goals. Allowing students to solve real world problems can be costly and risky. Observing students solving problems requires lots of instructors, and the observation process may interfere students. One-on-one coaching is not affordable when an instructor needs to teach a large class with students from diverse backgrounds.

Modern education technologists attempt to solve these problems using computer-based learning environments. Computer technology can simulate real world problems, allowing students to do endless practices with minimal costs and risks. Computers can also non-intrusively record all the actions (i.e., all