Chapter 1

Educational Data Mining: A Guide for Educational Researchers

Osman Kandara
Southern University and A&M College, USA

Eugene Kennedy
https://orcid.org/0000-0002-1174-2945
Louisiana State University, USA

ABSTRACT

This chapter presents a comprehensive discussion of educational data mining and its potential for educational research. The origins of data mining and the emergence of educational data mining are discussed. The variety of data generated in education (e.g., text, speech, performance, etc.) are described and the challenges of mining these data for useful information are identified. Techniques for mining these data are discussed. Software used to mine these data are noted and issues of theory and ethics are considered. Examples from published literature are cited throughout the chapter and recommendations for educational researchers are offered.

INTRODUCTION

In the past few decades there has been an unprecedented increase in data available to educational researchers (Aldowah, Al-Samarraie, & Fauzy, 2019). As schools, universities, and other educational institutions have increasingly incorporated technology (electronic databases, online courses, virtual tutors, etc.) in their operations and processes, large petabytes of data have been created (Urbina De la Calleja, 2017). These data have given researchers access to nearly every aspect of education, including near real-time insights into teaching and learning. As educators increasingly incorporate technology (e.g., video monitoring of classrooms, online course delivery, intelligent tutors, etc.) into their day-to-day interactions with students, researchers have greater capability to capture and examine learning, as it occurs, than ever before (see for example, Lynch, Merceron, Desmarais, & Nkambou, 2019). In the past researchers were often faced with the choice of collecting data at the beginning or end of some process or to expend

DOI: 10.4018/978-1-7998-1173-2.ch001
limited resources to make and process observations. Now, with web-based tools and technologies, text, voice and video mining, it is possible to capture and process data from a variety of formal and informal interactions and events, as they occur (Car, Sheikh, Wicks, & Williams, 2019). On a scale not seen in the past, the educational researcher can peer into ongoing educational processes and collect, analyze and present results on a range of cognitive, affective and behavioral data. Significantly, these observations are often unobtrusive and do not interrupt the natural flow of activity in social contexts, thus limiting some threats to internal validity such as occur when subjects present themselves in a positive light. Further, data mining as a research tool reverses the traditional paradigm in which theory leads to predictions which are confirmed or refuted by empirical data (Lukosius & Hyman, 2019). As a research strategy, data mining starts with discovery in data and then can be used to inform theory. In this chapter we present a review of the origins and nature of educational data mining and learning analytics, including examples of their use from published studies. Finally, software and other tools needed are discussed.

BACKGROUND

Data mining emerged in the 1990s in response to the proliferation of mega datasets as corporations, government agencies and other entities increasingly moved operational records and other data to electronic databases (Gandomi & Haider, 2015). These so-called “Big Data” were often linked allowing, for example, businesses to access a prospective customer’s entire history of purchases, employment, and the like (Chopra, Gautreau, Khan, Mirsafian, & Golab, 2018). Similarly, the emergence of social media platforms such as Facebook, normalization of email use, and other developments gave potential employers a history of interactions of prospective employers beyond anything possible in the past (Injadat, Salo, & Nassif, 2016). Finally, the ubiquitous mobile devices that are now part of everyday life provide an uninterrupted stream of data about social interactions and online activity for billions of people (Cheng, Fang, Hong, & Yang, 2017). Data mining grew out of the recognition that these databases provides invaluable information which can be meaningfully extracted, condensed and presented. As a discipline, data mining is concerned with the complexities of this process (Koedinger, D’Mello, McLaughlin, Pardos, & Rose, 2015; Wang et al., 2018).

The international association of data mining emerged from a special interest group of computer scientists and the business world (Gandomi & Haider, 2015). The primary focus and goals of this entity was reflected in its name, Special Interest Group for Knowledge Discovery in Databases (SIGKDD). Data mining, at its heart, is a process designed to uncover trends, associations, groupings, patterns, etc. in data that were not previously known. The input to a data mining process may consist of terabytes of data elements from widely discrepant sources and covering a range of characteristics. The goal of data mining is to scrub these data and find consistencies and regularities that can inform decision making, practice and perhaps promote greater understanding (Peña-Ayala, 2014). To accomplish this goal, data mining typically follows three steps: (a) Data Screening and Preparation, (b) Model Building and Validation, and (c) Deployment.