Chapter 12

Machine Learning Data Analysis for On-Line Education Environments: A Roadmap to System Design

Marian Cristian Mihaescu

University of Craiova, Romania

ABSTRACT

The increase of e-Learning resources such as interactive learning environments, learning management systems or intelligent tutoring systems has created huge repositories of educational data that can be explored. This increase generated the need of integrating machine learning methodologies into the currently existing e-Learning environments. The integration of such procedures focuses on working with a wide range of data analysis algorithms and their various implementations in form of tools or technologies. This paper aims to present a self-contained roadmap for practitioners who want to have basic knowledge about a core set of algorithms and who want to apply them on educational data. The background of this research domain is represented by state-of-the-art data analysis algorithms found in the areas of Machine Learning, Information Retrieval or Data Mining that are adapted to work on educational data. The main goal of the research efforts in the domain of Intelligent Data Analysis on Educational Data is to provide tools that enhance the quality of the on-line educational systems.

INTRODUCTION

Data Analysis for on-line educational environments as well as Educational Data Mining (EDM) may be applied to extract knowledge from educational assets. In this context educational assets may be represented by educational content such as documents, images, quizzes, etc. or activity data such number of sessions, time spent on line, number/dimension of sent or received messages, etc. The intelligent data analysis of this huge amount of data in an attempt of finding information, knowledge or even wisdom may become very useful for the e-Learning environment that manages and produces the analyzed data.

DOI: 10.4018/978-1-5225-3129-6.ch012
The main goal is to carry out information retrieval activities as an integrated functional part of the beneficiary e-Learning system.

It is a well-known fact that data understanding and pre-processing constitutes the main work of the data analysis and mining process. Although learning algorithms are not especially designed for data mining or information retrieval these three data analysis techniques share a lot of common fundamental approaches. Even if many e-Learning software systems do store huge amounts of usage and educational data and they are designed to support learning and teaching, they usually do not analyze the data they store. However, the field of educational data mining is emerging precisely because valuable information may be obtained from analyzing and mining the data stored by the educational software. As a consequence, the process of IR on educational data becomes sometimes long and tedious but the outcomes may be surprisingly useful.

Since its inception as a domain, educational data mining requires domain experts who use machine learning rather than machine learning experts who study education. The most commonly used methods are thus highly dependent on domain expertise. Examples include domain experts constructing data features, generating priors, and developing initial seed models. An expertise-based approach is highly effective for educational data, but a reliance on domain experts has risks: if the domain expert’s prior beliefs are wrong then the results will tend to be biased. The process can also be time-consuming and difficult for other researchers to replicate.

The chapter is self-contained and consists of six sections. The next section of the chapter presents the IR on educational data and EDM related works and state of the art achievements in the domain. The third section provides an introduction of the main classes of algorithms with an emphasis on the guidelines of how they may be integrated into IR systems. The fourth section presents the main tools and technologies that may be used for designing successful IR system that work with educational data. The final section provides the conclusions and the main promising roads in the domain of IR on educational data.

RELATED WORK AND CURRENT STATE OF THE ART RESEARCH

Applying advanced algorithms on educational data started back in 1995 as in Barnes (2005) or Choquet et al. (2009). At that moment, models that were used in psychometrics literature made a shift towards educational data mining. Right from the beginning EDM domain defined itself as an interdisciplinary field. Theoretical as well as practical aspects from machine learning, statistics, information retrieval and visualization, and computational modeling are usually found in EDM research area.

An important step forward was made by Romero and Ventura (2007) in the attempt to categorize the main directions of the work that needs to be performed in EDM. At that moment there were defined the main types of procedures that may be applied for solving different problems from EDM. The two categories of procedures were defined: web mining (i.e., clustering, classification, association rule mining, and text mining) and visualization. From this perspective it is obvious that a main source of data that was envisioned to be analyzed was represented by the logs (i.e., the activity) of student-computer interaction. On the other hand, IR on educational data tackles with other types of data: databases (i.e., classical IR), documents, images as in works of Stănescu et al. (2010).

Later, Baker (2010) presented a more refined taxonomy of research work in EDM. The main tracks were prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models. Among the subcategories there were set classification, regression, association rule min-