EDUCATIONAL DATA MINING

Oliver Scheuer
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken
Germany
Oliver.Scheuer@dfki.de

Bruce M. McLaren
Carnegie Mellon University
Pittsburgh, PA
U.S.A.
bmclaren@cs.cmu.edu

Synonyms

e-Learning data analysis, analysis of learning data, Education Analytics

Definition

Computer-based learning systems can now keep detailed logs of user-system interactions, including key clicks, eye-tracking, and video data, opening up new opportunities to study how students learn with technology. *Educational Data Mining* (EDM; Romero, Ventura, Pechenizkiy, & Baker, 2010) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data – patterns that would otherwise be hard or impossible to analyze due to the enormous volume of data they exist within. Data of interest is not restricted to interactions of individual students with an educational system (e.g., navigation behavior, input to quizzes and interactive exercises) but might also include data from collaborating students (e.g., text chat), administrative data (e.g., school, school district, teacher), and demographic data (e.g., gender, age, school grades). Data on student affect (e.g., motivation, emotional states) has also been a focus, which can be inferred from physiological sensors (e.g., facial expression, seat posture and perspiration). EDM uses methods and tools from the broader field of *Data Mining* (Witten & Frank, 2005), a sub-field of Computer Science and Artificial Intelligence that has been used for purposes as diverse as credit card fraud detection, analysis of gene sequences in bioinformatics, or the analysis of purchasing behaviors of customers. Distinguishing EDM features are its particular focus on educational data and problems, both theoretical (e.g., investigating a learning hypothesis) and practical (e.g., improving a learning tool). Furthermore, EDM makes a methodological contribution by developing and researching data mining techniques for educational applications. Typical steps in an EDM project include data acquisition, data preprocessing (e.g., data "cleaning"), data mining, and validation of results.

Background

Historically, EDM is a relatively new scientific discipline. Although researchers have been recording and analyzing data from educational software for a long time, only recently has EDM been established as a field in its own right, through conferences (Internal Conference on Educational Data Mining, started in 2008) and a scientific journal (Journal of Educational Data Mining (JEDM), first issue published 2009). EDM research is also presented and promoted at educational technology conferences, such as the International Conference on Artificial Intelligence in Education (AIED), the International Conference on Intelligent Tutoring Systems (ITS) and the International Conference on User Modeling, Adaptation, and Personalization (UMAP).

EDM borrows from and extends related fields such as Machine Learning (the study of computer programs that learn from and improve with empirical data), *text mining* (approaches to finding patterns in natural language text) and *statistics*. Other important influences are *psychometrics* (the study of psychological instruments to measure human skills and traits) and *web log analysis* (approaches to identify user profiles and navigational patterns of web site users).

EDM provides a rich toolbox of analysis techniques for a variety of problems in educational research and technology development (Romero et al. 2010; International Working Group on Educational Data Mining, n. d.):

- *Scientific inquiry and system evaluation*. EDM can contribute to the evaluation of learning systems and the development and testing of scientific theories on technology-enhanced learning (TEL). Exploratory analyses can be used to identify regular (or unusual) patterns in data, for instance, problem-solving strategies of students and patterns of successful and unsuccessful collaboration, thus helping to formulate new scientific hypotheses. EDM can be used to compare different interventions, for instance, how different types of practice compare to one another (e.g., in language learning, is it more efficient to reread the same stories or to read a variety of stories?). To simplify the execution of studies, EDM researchers developed computerized methods to randomize treatment assignment and to capture data. Finally, EDM researchers have developed new evaluation methods that are based on specific models of learning (e.g., learning curves and Bayesian Knowledge Tracing).
- *Determining student model parameters*. A *student model* is a data structure, typically used in Intelligent Tutoring Systems (ITS), that keeps track of relevant student characteristics over time (e.g., how a student's mastery of a particular skill improves with practice) through inferences from observable user actions (e.g., student answers to quiz questions). Student models allow systems to adapt to students and situations (e.g., selecting an exercise with an appropriate level of difficulty). Often, these inferences are based on parameterized probabilistic models. For instance, *Bayesian Knowledge Tracing* (Corbett & Anderson, 1995) – a modeling approach frequently used to implement Mastery Learning – uses parameters to represent the probability that a student only guessed the correct solution. System developers are confronted with the problem of how to choose appropriate parameter values. One solution is to use EDM to *estimate parameter values* from real data.
- *Informing domain models*. Student models are typically built upon a *domain model*, i.e., a model that formally describes the domain of instruction in terms of concepts, skills, learning items and their interrelationships. To enable student models to accurately predict knowledge and skills, it is important that the domain model reflects aspects of human cognition, such as knowledge and problem solving skills. EDM can help in the design, refinement and evaluation of domain models. For instance, EDM has been used to induce domain structures from data, to detect empirically plausible knowledge components that were originally not included in a domain model, and to compare skill models with different granularities (e.g., fine-grained models with many skills and coarse-grained ones with only a few skills).
- *Creating diagnostic models*. Learning systems typically diagnose student's progress to drive adaptation and feedback. Yet, the patterns these diagnoses are based on are often complex and/or not well understood. EDM provides tools to explore existing data to sharpen understanding of patterns of interest and to induce diagnostic models from data. For instance, models have been developed to identify gaming-the-system behavior (students exploit properties of the learning system to "succeed," e.g., by excessive use of a hint function to get the answers to problems), students' learning styles, off-

topic contributions in discussions, problems during collaboration, and students' emotions. Once such models have been developed, corresponding analysis results can be fed into a student model.

- *Creating reports and alerts for instructors, students and other stakeholders.* It is usually difficult for teachers in computer-based learning scenarios to monitor their students' learning progress and problems in real time (or even shortly after) because of a lack of face-to-face interaction, differences in time and place and the impracticality of immediate and rapid manual analysis of computer logs. EDM can be used to build teacher tools that employ statistics, visualizations and other ways of representing information in an intelligible way to facilitate the exploration of data, to increase awareness of students' current learning and to pinpoint possible problems that may require remediation. For instance, tools have been developed that provide statistics on content usage, student performance, and participation in collaborative activities, as well as visualizations of navigation paths and social networks. Similarly, such information can be provided to students to support awareness and metacognition.
- *Recommending resources and activities.* EDM can be used for adaptive instructional support by determining learning resources and activities that are appropriate with respect to a learner's needs, interests, preferences, skill level and past activities. For instance, a resource could be recommended that has been used by other students with a similar profile who successfully mastered an intended learning goal before.

EDM uses a wide range of methods to analyze data. The following taxonomy builds upon previously proposed taxonomies (Romero & Ventura, 2007; Baker & Yacef, 2009):

- *Supervised model induction* comprises machine learning techniques that infer prediction models from training instances for which the values of a target attribute are *known*. Prediction models accept instances as input (typically described as an attribute vector) and output a prediction for the target attribute. Models that predict categorical target values are called *classification* models; models that predict continuous target values are called *regression* models. Prediction models can be based on different representations, for instance, *Decision Trees*, *Support Vector Machines* (both classification) and *linear regression* model (regression). An example application in EDM is the categorization of discussion contributions into on-topic and off-topic contributions (the target attribute) based on the list of terms extracted from the contribution's text (the attribute vector).
- *Unsupervised model induction* comprises machine learning techniques that infer models from training instances for which the values of a target attribute are *not known*. Unsupervised methods use a bottom-up approach, that is, patterns and structures are searched in the input space without explicitly defined target categories or labeled examples. A widely used approach is *clustering*, which is used to identify groups of instances in a training set that are "similar" in some respect. Typically, some kind of distance measure (e.g., Euclidian distance) is used to decide how similar instances are. Once a set of clusters has been determined, new instances can be classified by determining the closest cluster. One well-known clustering algorithm is *k-means clustering*. Example applications in EDM are the identification of similar course materials or similar interaction sequences in collaborative learning.
- *Parameter estimation* comprises statistical techniques to infer parameters of probabilistic models from given data. These models can be used to predict the probability of events of interest. The approach is based on the assumption that the model has a given parametric form (e.g., a Gaussian distribution with the parameters *mean* and *variance*). An example application in EDM is the estimation of *Bayesian Knowledge Tracing* (BKT) parameters. BKT is used to determine the probability that a student has mastered a skill based on the history of past performances. A BKT model can be understood as a *Dynamic Bayesian Network* with four parameters (*prior*, *guess*, *slip* and *learn rate*). These parameters can be determined, for instance, with the *Expectation-Maximization* algorithm. Besides parametric estimation methods there are also *nonparametric methods* that do not assume a specific parametric form.

- *Relationship mining* is concerned with the identification of relationships between variables – relationships which might be associative, correlational, sequential or causal in nature. For instance, a common approach to association rule mining is to learn IF-THEN rules that exceed a minimum "support" and "confidence" threshold. S*upport* denotes the relative frequency of transactions that match both the IF and THEN part of the rule. C*onfidence* denotes the relative frequency of transactions that match the THEN part of the rule within the set of transaction that match the IF part. *Apriori* is, for instance, a classical association rule algorithm. An example application in EDM is the identification of mistakes that frequently occur together (e.g., students who made the mistakes *A* and *B* also often made mistake *C*).
- *Distillation of data for human judgment* aims at representing data in intelligible ways using statistics, visualization methods and interactive information interfaces. For instance, average performance scores can be computed for each student and presented to a teacher in ascending order in a bar chart. Another example is learning curves, which plot a student's performances (e.g., response time) against the number of opportunities of practicing a skill. An ideal learning curve shows that performance improves smoothly and monotonically, approximately following a power law or exponential function. On the other hand, learning curves with spikes indicate that another skill might interfere with the actually modeled skill, that is, the skill model could be improved.
- *Discovery with models* comprises approaches that bootstrap already existing models to make discoveries rather than computing new models from scratch. For instance, a prediction model could be applied to a data set to predict the values of a target category of interest. The predictions themselves could be used as data in other analyses again, for instance, they could be correlated with a target category of another prediction model. Another example is to scrutinize the different components of an existing prediction model to learn about factors that influence the prediction (e.g., how do interaction sequences of successful and failed knowledge sharing in collaborative learning differ?).

Important Scientific Research and Open Questions

A typical characteristic of educational data is its *non-independence*. For instance when we collect data from education discussions and want to classify whether contributions are on-topic or off-topic we have to consider that contributions are not statistically independent of one another since multiple contributions might stem from the same student or discussion. This might harm the computation of models (standard machine learning schemes typically have the built-in assumption of independent training examples) as well as the validation of models (e.g., a *cross-validation* might lead to biased results when training and test set are not independent).

Results from EDM research are typically achieved in the narrow context of specific research projects and educational settings (e.g., a particular school). The question arises how general such results are, for instance, whether the same student model parameters also can be used with other student populations, or whether a predictive model is still reliable when used in a different context. Therefore, there is an increasing need for *replication studies* to test for *broader generalizations*.

As a practical consequence of this need, EDM researchers have become increasingly more interested in *open data repositories* and *standard data formats* to promote the exchange of data and models. An example is the PSLC DataShop (http://pslcdatashop.org), a repository for educational data that has been opened to the EDM community, which provides data import and export facilities as well as analysis and visualization tools.

Cross-References

→ Advanced learning technologies

→ Artificial intelligence

→ Computer-enhanced learning and learning environments

→ Design of learning environments

→ Intelligent Tutorial Systems

→ Measurement of learning processes and outcomes

→ Probability theory in machine learning

→ Statistical learning technique

→ Supervised learning

→ Unsupervised learning

References

Baker, R. & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining (JEDM), 1*(1), 3–17.

Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

International Working Group on Educational Data Mining. (n. d.). *Educational Data Mining Home Page.* http://www.educationaldatamining.org. Accessed 23 December 2010.

Romero, C., & Ventura, S. (2007). Educational Data Mining. A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.d. (Eds.). (2010). *Handbook of Educational Data Mining.* CRC Press.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd edition). San Francisco: Morgan Kaufmann.