Educational Data Mining for Teaching and Learning

Zhi-Jun PEI\textsuperscript{1,a}

\textsuperscript{1}School of Electronic Engineering, Tianjin University of Technology and Education, China
\textsuperscript{a}peizj@tute.edu.cn

Keywords: Educational data mining, Teaching and learning, Prediction, Relationship mining, Structure discovery.

Abstract. Educational Data Mining (EDM) seeks to use plethora of data in educational context to better understand learners and learning. It also concerns about computational approaches that combine data and theory to transform educational practice. The key EDM techniques for teaching and learning are first discussed, which include prediction models, structure discovery, and relationship mining. Then the most typical tasks related to teaching and learning that have been resolved through EDM techniques are reviewed, such as feedback for supporting teachers, recommendations for students, prediction of student performance, and detection of undesirable student behaviors.

Introduction

It is well known that Data Mining (DM) has been used by businesses, scientists and governments to extract useful information from volumes of data [1]. Education domain has always had the power to generate a large amount of data. As new learning technologies continue to penetrate all facets of education, a plethora of useful data traces are generated, which provides a gold mine of educational data [2]. For example, education is increasingly occurring online or in educational software, resulting in an explosion of data that can be used to improve educational effectiveness and support basic research on learning. Educational Data Mining (EDM) is just the application of DM techniques to educational data. And its objective is to analyze these type of data in order to resolve educational research issues [3], such as to better understand students and the settings in which they learn. EDM has emerged as a research area in recent years for researchers all over the world from related research areas. From a practical view point, EDM allows to discover new knowledge, based on usage data of students, in order to help to potentially improve the quality of education and to obtain a more effective learning process [4]. However, compared with the successfully application of DM in e-commerce systems, there has been comparatively less progress in education, although there is currently an increasing interest in applying data mining to the educational environment [5].

The EDM process converts raw data coming from educational systems into useful information, which could potentially have a great impact on educational research and practice. Just as other application areas of data mining like business, this process follows the same steps as the general data mining process, which consists of pre-processing, data mining and post-processing, as shown in figure 1. However, educational data and problems have some special characteristics that require the issue of mining to be treated in a different way. Although most of the traditional DM techniques can be applied directly, others have to be adapted to the specific educational problem at hand. Furthermore, specific data mining techniques can be used for specific educational problems. Thus EDM not only use typical DM techniques such as classification, clustering, but also other approaches such as regression, visualization.

Figure 1. EDM Process.
Nowadays, there is a great variety of educational environments. And different data provided by those educational environments enable different problems to be resolved with data mining techniques. The discussion will focus on the EDM for teaching and learning in this paper.

**EDM Techniques**

EDM may be seen either as a research community or as an area of scientific inquiry. On one hand, EDM is a sister community to learning analytics [6]. On the other hand, EDM is concerned with the analysis of large scale educational data. We will focus on the area of scientific inquiry of EDM, especially in teaching and learning. Although a wide range of EDM methods have emerged, only some key EDM techniques, related to the teaching and learning, are provided here, including prediction, structure discovery, and relationship mining.

**Prediction**

The goal of prediction is to develop a model which can infer a single aspect of the data from some combination of other aspects of the data. Labels are required by the prediction for the output variable for a limited dataset, where a label represents some trusted ground truth information about the predicted variables value in specific cases. Ground truth can come from a variety of sources, such as state standardized exam scores. Prediction models are then can be used to predict what a value will be in contexts where its label may not to be directly obtained. In EDM, classifiers and regressors are the most common types of prediction models. Both have rich history in data mining and artificial intelligence, and are leveraged by EDM research.

Classification is firstly considered. In classifiers, the predicted variable can be either a binary or categorical variable. Some popular classification methods in educational domains include decision trees, random forests, decision rules, step regression, and logistic regression. In EDM, classifiers are typically validated through cross validation. Cross validation should be conducted at multiple levels. For instance, it is typically standard in EDM for researchers to cross validate at the student level in order to ensure that the model will work for new students. Regression is then considered. In regression, the predicted variable is a continuous variable. The most popular regressors within EDM is linear regression. A model produced through this method is mathematically same as the linear regression used in statistical significance testing. But the method for selecting and validating the model in EDM usage of linear regression is quite different than in statistical significance testing. Regressors can be validated through the same techniques as in classifiers.

**Structure Discovery**

Structure discovery algorithms attempt to find structure in the data without any ground truth or a priori idea of what should be found. In this way, this type of data mining contrasts strongly with prediction models, where ground truth labels must be applied to a subset of the data before model development. Common structure discovery algorithms in educational data include clustering, factor analysis, and domain structure discovery algorithms. Clustering and factor analysis have been used since the early days of the field of statistics, and were explored further by the data mining. And domain structure discovery emerged from the field of educational measurement.

As for clustering, its goal is to find data points that naturally group together, splitting the full dataset into a set of clusters. Clustering is particularly useful in cases where the most common categories within the dataset are not known in advance. If a set of clusters is optimal, each data point in a cluster will in general be more similar to the other data points in that cluster than the data points in other clusters. Clusters can be created at several different grain sizes. For example, schools could be clustered together to investigate similarities and differences among schools, students could be clustered together to investigate similarities and differences among students, or student actions could be clustered together to investigate patterns of behavior. Clustering algorithms typically split into two categories. Hierarchical approaches assume that clusters themselves cluster together. And
non-hierarchical approaches assume that clusters are separate from each other, such as k-means, Gaussian mixture modeling, and spectral clustering. As for factor analysis, the goal is to find variables that naturally group together, splitting the set of variables into a set of latent or not directly observable factors. In EDM, factor analysis is general used for dimensionality reduction, which may be included in preprocessing to determine features. Factor analysis includes principal component analysis and exponential family principal components analysis.

**Relationship Mining**

In relationship mining, the goal is to discover relationships between variables in a dataset with a large number of variables. This may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover which relationships between any two variables are strongest. There are four types of relationship mining commonly used in EDM, including association rule mining, sequential pattern mining, correlation mining, and causal data mining. 

Association rule mining comes from the field of data mining, in particular from market basket analysis used in mining of business data. Sequential pattern mining also comes from data mining, with some variants emerging from the bioinformatics community. Correlation mining has been a practice in statistics for some time. Causal data mining also comes from the intersection of statistics and datamining.

First of all, the goal association rule mining is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value. For example, a rule might be found of the form:

**IF** student is frustrated **OR** has a stronger goal of learning than performance **THEN** the student frequently asks for help

Then, the goal of sequential pattern mining is to find temporal associations between events. Sequential pattern mining algorithms depend on a number of parameters to select which rules are worth outputting. These methods, like association rule mining, have been used for a variety of applications, such as to study what paths in student collaboration behaviors lead to a more successful eventual group project. And then, the goal of correlation mining is to find positive or negative linear correlations between variables. Correlation mining has been used to study the relationship between student attitudes and help-seeking behaviors. At last, the goal of causal data mining is to find whether one event or observed construct was the cause of another event or observed construct. Causal data mining is distinguished from prediction, which attempts to find actual causal relationships by looking at the patterns of covariance between those variables and other variables in the dataset.

**Enhancing Teaching and Learning**

Although many possible areas of application for EDM has been suggested [7], here we just focus on improvements of teaching and learning with educational data mining techniques, just discussed above.

**Feedback for Supporting Teachers**

It is possible to provide feedback with EDM methods to support course teachers in decision making, in order to take appropriate proactive action for the educational tasks, such as about how to improve students learning or organize instructional resources more efficiently. Providing feedback could extract completely new, hidden and interesting information in data. This is different from the tasks of data analyzing and visualizing, which only provide basic information directly from data.

The most common EDM techniques adopted in this task may be the association rule mining, which reveals interesting relationships among variables in large databases. Association rule mining has been used to solve the problem of feedback in the educational process. For example[8], it has been used to determine the relationship between each learning behavior pattern so that the teacher can promote collaborative learning behavior on the web; to find embedded information, which can be provided to teachers to further analyze, refine or reorganize teaching materials and tests in adaptive learning.
environments; to discover interesting associations between student attributes, problem attributes and solution strategies in order to improve online education systems for both teachers and students; and to improve an adaptive course design in order to show recommendations on how to enhance the course structure and contents. Several data mining models can also be applied together to provide feedback. For example, association rules, clustering, classification, sequential pattern analysis, and prediction have been used to enhance web-based learning environments to improve the degree to which the educator can evaluate the learning process; and association analysis, and clustering analysis have also been used to organize course material and assign homework at different levels of difficulty.

Data may come specifically from tests, questions, assessments. In this case of special type of feedback, the objective is improve the questionnaires and to answer questions such as what questions test the same information and which are of the most use for predicting course or test results. For example, several EDM approaches, including statistic correlation analysis, fuzzy clustering analysis, k-means clustering and fuzzy association rule mining, have been applied to support mobile formative assessment, in order to help teachers understand the main influencing factors in learner performance. Another special type of feedback involves the use of text data. In this case, the goal of applying EDM to educational text data is to summarize the learner discussion process, in order to provide instructor feedback.

**Recommendation for Students**

EDM approaches could be used to make recommendations directly to the students with respect to their personalized activities, the next task or problem to be done, which adapt learning contents, and interfaces to each particular student. The most common EDM techniques used for this task are sequential pattern mining, association rule mining, and clustering. For example, Sequential pattern mining has been developed for recommending lessons that a student should study next while using an adaptive hypermedia system, and for adapting learning resource sequencing; association rule mining has been used to provide students with personalized learning suggestions by analyzing their test results and test related concepts, and for course recommendation with respect to optimal elective courses; clustering has been developed for providing personalized course material recommendations based on learner ability, and to recommend to students those resources they have not yet visited but would find most helpful. To provide adaptive and personalized learning support, other EDM techniques may be employed, such as neural networks, and decision tree analysis.

**Student Performance Prediction**

The goal of student performance prediction is to estimate the unknown value of a variable that describes the student. In education, the predicted values are normally performance, knowledge, score or mark. This value can be continuous value for regression task or discrete value for classification task. Just as mentioned above, regression analysis finds the relationship between a dependent variable and one or more independent variables. And classification is a procedure in which individual items are placed into groups, based on a training set of previously labeled items. Prediction of a student performance is one of the oldest and most popular applications of EDM in education. Different techniques and models have been applied, such as neural networks, Bayesian networks, regression, and correlation analysis. For example, different types of neural network models have been used to predict the number of errors a student will make, or to predict performance from test scores; Bayesian networks have been used to predict a future graduate cumulative Grade Point Average based on applicant background at the time of admission, or to predict end-of-year exam performance through student activity with online tutors; several regression techniques have been used to predict student academic performance using step wise linear regression, or to predict end-of-year accountability assessment scores using linear regression; correlation analyses have been applied to predict a student final exam score in online tutoring.

**Undesirable Student Behaviors Detection**

Detecting undesirable student behavior is to discover those students who have some type of problem or unusual behavior such as erroneous actions, low motivation, playing games, misuse, cheating,
dropping out, or academic failure. Classification and clustering are mainly EDM techniques which have been used to reveal these types of students, in order to provide them with appropriate help in plenty of time. For example [8], classification algorithms have been used to detect problematic student behavior for preventing student dropout using decision tree neural networks, logistic regression and support vector machines, or to identify students with little motivation using decision trees, or to predict, understand and prevent academic failure among university students using decision tree algorithm and clustering algorithm. Other EDM techniques may also be used to carry this task, such as association rule mining for selecting weak students for remedial classes or to construct concept-effect relationships for diagnosing student learning problems, Bayesian networks to predict the need for help in an interactive learning environment, and stepwise regression to detect misplay and look for sources of error in the prediction of student test scores.

Summary
EDM has emerged as an up and coming research area related to the well known areas of data mining research. It could be used to solve very kinds of education tasks such as teaching and learning. The applications of EDM in educational environment are growing fast with a large number of specific tools specially developed for applying data mining algorithms in educational data. EDM uses computational approaches to analyze educational data in order to study educational questions. As for teaching and learning, the key EDM methods include prediction models, structure discovery, and relationship mining. However, it is not yet a mature area. And instead of the current individual proposals, it is necessary for researchers to develop more unified and collaborative studies.

Acknowledgement
This research was financially supported by Research Foundation of Tianjin University of Technology and Education under Grant No. KJY1312.

References