Educ Inf Technol
DOI 10.1007/s10639-017-9616-z

Educational data mining applications and tasks: A survey of the last 10 years

Behdad Bakhshinategh[1] **· Osmar R. Zaiane**[2] **·**
Samira ElAtia[3] **· Donald Ipperciel**[4]

Received: 14 November 2016 / Accepted: 19 May 2017

Abstract Educational Data Mining (EDM) is the field of using data mining techniques in educational environments. There exist various methods and applications in EDM which can follow both applied research objectives such as improving and enhancing learning quality, as well as pure research objectives, which tend to improve our understanding of the learning process. In this study we have studied various tasks and applications existing in the field of EDM and categorized them based on their purposes. We have compared our study with other existing surveys about EDM and reported a taxonomy of task.

Keywords Educational data mining · Surveys · Taxonomy of applications

✉ Samira ElAtia
selatia@ualberta.ca

Behdad Bakhshinategh
bakhshin@ualberta.ca

Osmar R. Zaiane
zaiane@cs.ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[2] Department of Computing Science, University of Alberta, Alberta, Canada

[3] Faculté Saint-Jean, The University of Alberta, Edmonton, Alberta, Canada

[4] Glendon College York University, Toronto, Ontario, Canada

Published online: 03 July 2017

1 Introduction

Educational Data Mining (EDM) is the field of using data mining techniques in educational environments. Currently there are many computer-based learning systems which gather large amount of usage data such as Learning Management Systems (LMS), Massive Open Online Courses(MOOCs) and Intelligent Tutoring Systems (ITS). Stone et al. define an LMS as "a centralized web based information systems where the learning content is managed and learning activities are organized. LMS represents a more general term for a technology framework that supports all aspects of formal and informal learning processes, including learning management, content management, course management, etc" (Wang 2014). It consolidates in one platform a number of functionalities, including "personal communication (email and instant messages), group communication (chatting and forums), content posting (syllabus, papers, presentations, lesson summaries), performance evaluation (question and answer repositories, self-assessment tests, assignments, quizzes and exams), and instruction management (message and grade posting, surveys, and online office hours)" (Naveh et al. 2012) while serving as point of departure to the entire web. The data collected by LMSs provide us with the opportunity of using data mining.

There exist various methods and applications in EDM. These applications can follow both applied research objectives such as improving and enhancing learning quality,as well as pure research objectives, which tend to improve our understanding of the learning process. Aside from the classification of applications based on their objectives, which is the focus of this study, EDM applications can also be categorized based on the targeted end user. The applications of EDM can target any of the stakeholders involved in educational systems, such as learners, educators, administrators and researchers themselves. Providing feedback, personalization and recommendations can improve the learning process of students. Making discoveries and providing decision support systems can help the educators by improving teaching performance and making decisions. Administrators are provided resources and tools for making decisions and organizing the institutions. Also, discoveries in the education field can help researchers have a better understanding of educational structures and the evaluation of learning effectiveness.

In this article, we will look into various possible applications of EDM and cognate methods that can be used to meet the needs of these applications. We focus on the end objectives of these applications and seek to group the applications in categories with similar purposes. Although we try to draw a line between different categories of applications, it should be noted that in some cases, there is no clear boundary between the applications. Some research may belong to more than just one of these categories. In some other cases, an application can be used as a tool for reaching another application. There are many such examples; for instance, creating reports of students' expected performance for educators. In this case, the end objective is providing reports which need some visualization techniques. However, prediction of student performance can be described as another application in EDM, which is needed prior to providing a report.

The interest of this paper is to present a synchronic landscape of emerging possibilities afforded by EDM. Great strides have been made in this field in the past few years, but with little awareness from what should be its primary intended audience, i.e. educators. This paper intends to remedy this situation. It is an opportunity to take stock of the current state of EDM and assess its scope. Similar reviews have been published in the past (see literature review, below), but given the speed with which the field is evolving, it is worth taking a new look and propose and updated classification of current practices. Because this paper is also intended for educators wanting to acquaint themselves with EDM, it makes sense to first present a brief historical synopsis of EDM, which contextualizes the current state of the field. We then discuss the many methods and applications of EDM, with a view to organizing them into coherent sets of activities. Finally, in order to highlight the contribution of the present paper, we present a literature review of the major works on EDM.

1.1 A brief historical synopsis

EDM bridges between two disciplines: education on the one hand,computing sciences on the other, where both data mining and machine learning as subfields of computing sciences are the focus. In EDM, these two fields became intertwined through the years, and it is very important to keep the focus on how the two have contributed thus far to advancements in both educational research and learning/teaching. Yet, alone, i.e. uninformed by educational theory, CS and data cannot advance much nor lead to a quality education as a social science. To this end, it is essential to keep both anchored into each other's advances, and that is one of the most important contribution of EDM to education.

The use of computers and CS in education is by no means recent. It dates back to the mid-20th century. Computers were then used as a tool to teach directly in a drill-based approach (Bates 2015). Certainly, this was at the height of behaviorism, and computers lent themselves readily to this approach. As Bates puts is:

"B.F. Skinner started experimenting with teaching machines in 1954, based on the theory of behaviourism. In essence programmed learning structures information, provides immediate feedback to learners, and tests learning. This use of machines based on a behaviourist approach was called computer-assisted learning (CAL) or computer-based training (CBT), but went out of fashion in the 1980s, mainly because it did not handle well the higher levels of learning such as critical thinking, analysis and synthesis that are required at a university level, although CBT is still used in training in the workplace. (Bates 2015, p.9)"

The operative word in this initial use of computers in education is perhaps "training", as it underscores the repetitive mechanical way in which the first three levels of Bloom's taxonomy of the cognitive domain are targeted; namely knowledge, comprehension, application. At the time, targeting the higher level thinking skills was not in the picture. Yet, CBT is still in use today and enjoys continued success when the purpose is training individuals to perfect a specific task.

The first experimentation using online discussion forums was attributed to Murray Turoff and Roxanne Hiltz in the second half of the 1970s (Bates 2015). They coined the term "computer-mediated communication." By the late 1980s, many scholars around the world were experimenting with the use of computers in education in ways that could be classified into two categories: either with a focus on the use of computers “for automated or programmed learning” or on the use of computers for communication between students and instructors, and among students (ibid).

Up to this point, isolated computers were being used. Collecting, generating and analyzing data was an arduous task. Then, a revolution occurred after 1991, as the World Wide Web was formally launched. The first learning management systems were developed in 1995 (WebCT). Online teaching and e-learning environments became a reality. The time-consuming methods to load and search for materials was dramatically reduced. We moved to a more complex and nuanced use of computers for learning that transcended mere “training.”

The first time the word Education Data Mining was used dates back to 2005 (Romero and Ventura 2007). This occurred during the 2005 annual conference of the Association for the Advancement of Artificial Intelligence (AAAI’05) in Pittsburgh, USA, in a workshop on Educational Data Mining (Romero and Ventura 2007).

Looking at the paper presented during this workshop, two things stand out. First, it was all about the technical aspects of collecting and analyzing data. Second, the focus was on computer education and computing sciences training. In 2009, a journal and an international conference on EDM were established (Baker and Yacef paper). Since then, the field has taken off. 10 years after that first workshop, we revisit the field and present a survey of the use and progress of EDM in higher education teaching and research.

2 Methods and applications

The methods of EDM are the same as those in the data mining field in general. There are multiple methods to use in EDM for each of the various applications. Among these methods, the most used are classification and regression (1), clustering (2), association rule mining (3), discovery with models (4), outlier detection (5), social network analysis (6), text mining (7), sequential pattern mining (8) and visualization techniques (9) also referred to as distillation of data for human judgment. Based on the research (Romero and Ventura 2010) done by Romero and Ventura in 2009, the most commonly applied data mining tasks are regression, clustering, classification and association rule mining.

Applications and tasks in EDM can be categorized based on different properties. Multiple surveys of EDM exist, which have listed possible applications of EDM. We will look into these surveys in more detail in the literature review section. Considering these surveys and reading into the research examples they have provided, as well as the recent studies published in journals of educational data mining, we propose a new list of EDM categories. In this list, we have tried to consider all the categories mentioned in previous surveys and in the literature, as well as new categories which we think need to be added. These new categories of applications can be explained by the growth of interest in EDM. In previous studies, possible applications of EDM have been introduced sometimes in no specific order, sometimes based on the number of research papers done in each of the categories. We try to group possible applications of EDM into categories based on their end objective. We have tried to group different applications together as much as possible to better highlight the similarities and differences.

We have identified 13 categories of applications, as shown in Fig. 1, forming a new taxonomy tailored specifically to EDM, thus setting EDM as a specific subfield of data mining. Four applications are grouped under "Student Modeling", six under "Decision Support Systems" and the last three are presented as "Other" because they differ from the other applications.

In the remainder of this section we will describe and illustrate these applications with the help of research examples related to each category of applications for more clarity.

3 Student modeling

Student modelling is a process devoted to representing cognitive aspects of student activities,such as analyzing the student's performance or behaviour, isolating underlying misconceptions, representing students' goals and plans, identifying prior and acquired knowledge, maintaining an episodic memory and describing personality characteristics (Self n.d.; Chrysafiadi and Virvou 2013).

We have used this definition as a guide for categorizing some of the applications in EDM. All applications in this category present a model that describes

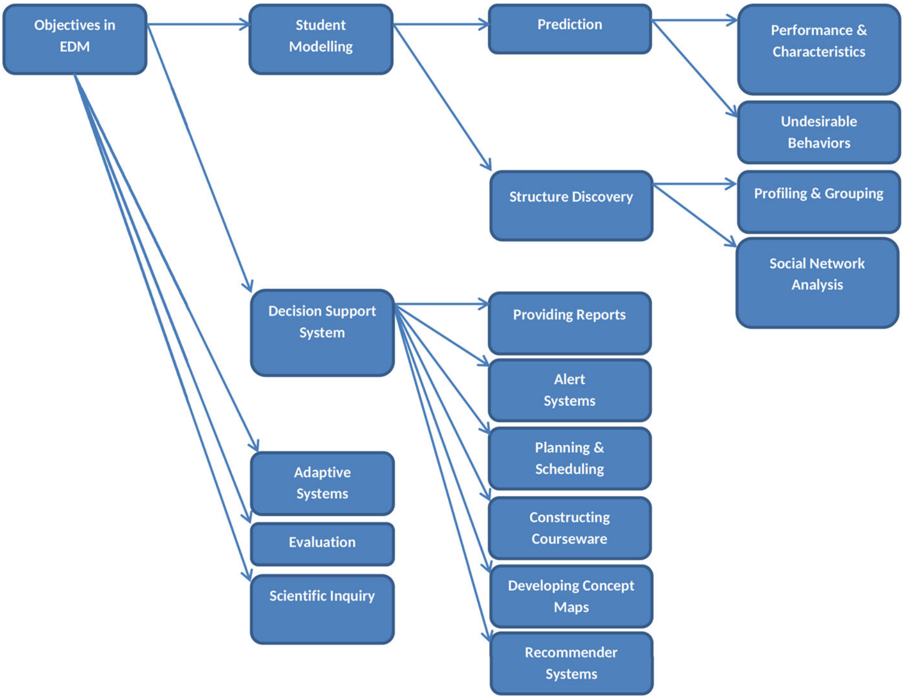

Fig. 1 Taxonomy of Applications in EDM

students attempting to reach their objective. Based on the literature review in 2013 (Chrysafiadi and Virvou 2013), there are different characteristics in student modelling, namely, knowledge and skills (1), errors and misconceptions (2), learning styles and preferences (3), affective and cognitive factors (4) and meta-cognitive factors (5). Modelling student activities and behaviour can be used for predicting values representing students (the characteristics above) or discovering structures that describe students. As a result, there are two sub-categories in student modelling: prediction and structure discovery. In prediction, we usually know a specific attribute that we wish to predict and in structure discovery, we may not know the specific attribute or it may be only defined as a structure, instead of a single property. It is also important to mention that there might not be a clear line between these two subcategories in all cases; but as there are enough differences in the objective of these two groups, it seems preferable to distinguish them.

3.1 Predicting student performance, achievement of learning outcomes or characteristics

In this set of applications/tasks, the objective is to estimate a value or variable describing students. This value can indicate students' performance, achievement of learning outcomes or characteristic. Most of the existing publications are focused on prediction of students' academic performance, but studies also look into characteristic such as collaboration with other students. The most widely used methods for predicting student performance and characteristics are regression and classification but other techniques have also been used such as clustering and feature selection. Miller et al. have used Lasso feature selection for identifying student characteristics which impact the learning. They compare the DM method with other models and examine if the selected features can be used for predicting student performance (Miller et al. 2015). Zimmermann et al. introduced a model-based approach to predict graduate-level performance using indicators of undergraduate-level performance. Feature selection and prediction techniques have been used in this study (Zimmermann et al. 2015). Galyardt and Goldin have used the recent student usage data in order to improve the prediction accuracy of the system in ITS (Galyardt and Goldin 2015). Research by Waters et al. identifies the collaboration of students in online courses using Bayesian classification (Waters et al. 2014). Sabourin et al. develop a model that identifies engagement of students based on their off-task behaviour in educational software which investigates whether or not off-task behaviour can be a self-regulator of emotions (Sabourin et al. 2013). Cocea and Weibelzahl estimate the motivational level of learners using decision trees (Cocea and Weibelzahl 2006).

3.2 Detecting undesirable student behaviors

This set of applications/tasks are similar to the prediction of student performance and characteristics, but in this category, the focus is mainly on detecting undesirable student behaviour, such as low motivation, erroneous actions, cheating, dropping out, academic failure, etc. The main data mining methods used in this category of applications are classification and clustering but other techniques are also applicable, such as feature selection and outlier detection. An example of this group of applications is the research done by Bravo and Ortigosa in which they propose an approach for detecting potential symptoms of low performance in e-learning using production rules (Bravo and Ortigosa 2009). In another study, Dekker et al. have used a decision tree classifier to predict student drop in an electrical engineering program (Dekker et al. 2009). Lykourentzou et al. used multiple machine learning techniques such as Support Vector Models and neural networks for students drop out prediction (Lykourentzou et al. 2009).

3.3 Profiling and grouping students

As the title of this category implies, the objective in this set of applications/tasks is to profile students based on different variables such as characteristics and knowledge, or using these information to group students for various purposes. Grouping students can be done based on various properties of profile information. This task is often different from clustering similar students with each other, as the purpose is to group students so as to complement each other. Also, when clustering students, one is looking for the greatest dissimilarity between clusters, but this may not be the case in grouping tasks. When using a grouping task for forming teams in a course project, one prefers to have groups that are similar, while comprising dissimilar students that can complement each other. In a way similar to other categories of applications, different data mining methods can be used for these tasks, such as feature selection and clustering. As examples of this category, Azarnoush et al. proposed a method for learner segmentation using a dissimilarity measure based on a random forest (Azarnoush et al. 2013); Kinnebrew et al. used sequence mining techniques to identify learning behaviour patterns differentiating distinct groups of students (Kinnebrew et al. 2013). Harley et al. studied the task of clustering and profiling of learners based on their interactions with an intelligent tutoring system (Harley et al. 2013).

3.4 Social network analysis

In this category of applications, the purpose is to obtain a model of students in the form of a graph, showing different possible relationships among them. In other applications of modelling, the focus is mostly on individuals, but in social network analysis (SNA), the focus is on the relationships between individuals. As an example, collaboration is a property assigned to the relationship between individuals, and to study it, one must model the relationships as well as the individuals. Rallo et al. used data mining and social network analysis to model the dynamics and structure of educative online communities (Rallo et al. 1999). Reffay and Chanier used social network analysis to measure cohesion in collaborative distance learning environments (Reffay and Chanier 2003). Reyes and Tchounikine studied structural properties of learning groups based on a relational perspective using social network analysis techniques (Reyes and Tchounikine 2005).

4 Decision support systems

The other major group of applications/tasks in EDM is the decision support systems. Applications devoted to this category enhance the process of learning by helping stakeholders make decisions. Examples of this category are: providing feedback, creating alerts, planning, generating recommendations and enhancing the courseware. The target of these decision support systems is mostly the instructor, but it can also be the student, administrators or researchers.

4.1 Providing reports

Data analysis and visualization can be used as one part of many other applications, but it can also be an application by itself in educational environments by providing useful information to educators and administrators to help them with decision making. As a result, the purpose of this category of applications is to find and highlight the information related to course activities which may be of use to educators and administrators and provide them with feedback. The results of most of the applications grouped in "Student modelling" can be used for creating reports. Examples of this are: providing feedback on student performance or characteristics, describing the connections and collaborations through social network analysis and creating reports from the profile information extracted with the help of profiling methods. An example of this category of applications is the research done by Romero et al. in which they used association rule mining to provide feedback to instructors from the multiple-choice quiz data (Romero et al. 2013).

4.2 Creating alerts for stakeholders

This category of applications is similar to applications in the student modelling category. Usually, the purpose is to predict student characteristics and detect unwanted behaviour, but it mainly serves as an online tool for informing stakeholders or creating alerts in real time. Examples of situations in which alerts may be needed are in cases of low motivation, misuse, cheating, etc. An example of study in this category is the research of Knowles, which introduces a dropout early warning system using statistical models and regression (Knowles 2015). In another study, Macfadyen and Dawson have developed an early warning system for educators using performance prediction (Macfadyen and Dawson 2010).

4.3 Planning and scheduling

The objective of this category of applications is to help stakeholders with the task of planning and scheduling. It can help educators and administrators with planning future courses or recourse allocation, assisting in the admission and counselling processes or any other tasks involved in planning and scheduling (Romero and Ventura 2010). It can also help students with course enrollment planning, in which case it has some common ground with recommender systems. In research with the objective of planning and scheduling, various methods have been used, such as discovery with models, cluster analysis and classification. Hsia et al. enhanced course planning by establishing the probability of enrollees completing their courses based on the student's preferences and profession (Hsia et al. 2008). The research of Delavari et al. discovers explicit knowledge that can be useful for decision making processes as well as proposing analytical guidelines for higher education institutions (Delavari et al. 2008). Huang et al. used cluster analysis, decision tree algorithm and neural networks for planning courses (Huang et al. 2009).

4.4 Creating courseware

Courseware is known as educational software providing content, videos, tests and other learning materials. In this category of applications, the objective is to help the educator create or development course material automatically using student usage information. An example of this category can be found in the research of García et al., in which they propose a system for developing, improving and maintening web-based courses using association rule mining and collaborative filtering (García et al. 2009).

4.5 Developing concept maps

Concept maps are "graphical tools for organizing and representing knowledge" (Novak and Cañas 2008). In this category of applications, the objective is to develop concept maps of various aspects to help educators define the process of education. In other words, concept maps provide domain models to educators. They can help with mapping different concepts to each other (i.e. ascertaining relationships). Examples of concept maps are: hierarchy of topics in course material, relationships of skills and test items, correlation of test items and knowledge components, etc.The research of Agrawal et al. presents a study navigator for studying electronic textbooks by creating a reference for concepts which students are reading (Agrawal et al. 2014). As a further example, Lee et al. used an Apriori algorithm to develop an automatically constructed concept map of learning, provided to educators (Lee et al. 2009).

4.6 Generating recommendation

In most of the surveys and books on EDM applications, generating recommendation is presented mainly as making recommendations to students. But recommendations can be targeted to any stakeholders. Examples of this category of applications are: course recommendations to students or test item recommendations to educators. The most common methods in recommender systems are collaborative filtering, content-based methods, association-rule based algorithms and hybrid approaches also used in EDM. Another method of generating recommendations is using discovery with models. For example, Vialardi et al. used a performance predictor model for generating recommendations (Vialardi et al. 2009). The predictor model predicts the success of each student in each course and will recommend courses which the student is most likely to be successful in. In another study, O'Mahony and Smyth develop a course recommender system using collaborative filtering (O'Mahony and Smyth 2007).

5 Other applications

5.1 Adaptive systems

This category of applications is related to the use of intelligent systems in computer-based learning, in which we need the system to adapt to the user's behaviour (this also referred to as "personalization"). This application is important, because in many online learning systems, numerous learners with different needs are involved with the system. And, as the number of participants grows, it becomes harder to meet the specific needs of all learners. Adaptive systems can help us meet the needs of every individual learner. This adaptation can take on the form of adapting the course material, instruction pace, providing hints, ordering and generating tests, etc. As an example, the research of Alaofi et al. explores the personalization of a digital library using the student's profile information in order to improve search results (Alaofi and Rumantir 2015). In another example, Tang et al. propose a method for personalizing courseware construction using data mining techniques for distance learning environments (Tang et al. 2000).

5.2 Evaluation

Evaluation is one of the aspects of the educational environment which may not always be intuitive, especially in computer-based learning environments. As an example, evaluation in ill-defined domains has been researched in intelligent tutoring systems and is known as a domain that lacks a definitive solution or the solution is dependent on the problem's conception (Lynch et al. 2006). As a result, the evaluation in these domains is challenging. In this category of EDM applications, the objective is to provide an evaluator to help the educators. This can be done in exploratory learning environments and computer-based courses. An example of these applications is the research of Mallavarapu et al., which proposes a computational method to measure and track students' spatial reasoning in an open-ended simulation (Mallavarapu et al. 2015). As another example, Hao et al. proposed a new method for scoring a game/scenario-based task using distance function (Hao et al. 2015).

5.3 Scientific inquiry

Similar to other fields of study, theories and hypotheses on the process of learning and possible improvements are used in education. One use of educational data mining can be testing or even developing theories based on the various records in big datasets. This category of applications mostly targets the researchers as the end user, but any of the developed or tested theories can be used in other applications targeting other stakeholders later.

6 End users in EDM tasks

For better clarification of the identified applications, we can look into the target users of each application. This has the added value of also showing the possible applications for the end users which have not been targeted yet. The end users in educational environments are learners (students), educators (instructors), administrators and researchers.

Learners have been the target of EDM in various applications such as grouping students, generating recommendations and adaptive systems. One important goal of EDM as a whole is improving the quality of learning; and in the process of learning, two groups of users come to mind first, i.e. learners and educators. Most of the applications in categories of student modelling and decision support systems target educators as their end users. Student modelling provides a better understanding of students' state of learning and decision support systems can directly help educators make better decisions for improving the learning process. This also applies to the administrator of educational institutions making higher level decisions. Researchers also represent a group of end users, as the objective of the research is to understand the learning process, develop theories and test them. As an example, researchers can use social network analysis (SNA) to pinpoint the properties that are valuable in prediction of student performance. Table 1 presents possible targeted users of each application. It is important to mention that any research in EDM may address one or more than one of these classes at once.

7 Review of previous surveys

Data mining has been used for making discoveries in educational environments for the last few decades. In the last decade, the availability of online datasets and more uses of online learning systems have garnered more attention to this field. The Journal of Educational Data Mining (JEDM), created in 2009, has brought researchers of this field together. The publication of two books in recent years shows the growth of interested in this field, *Handbook of Educational Data Mining* published in 2010 and *Educational Data Mining: Applications and Trends* in 2013. There has also been multiple survey articles published about EDM so far. In this paper, we have tried to integrate the ideas put forward in each of the surveys and books. In this section, we summarize the surveys and books related to applications of EDM, and the problems which have been addressed in each.

In the first survey published in JEDM and written by Baker and Yacef (2009), four areas of application have been mentioned, i.e. improving student models (1), improving domain models (2), studying the pedagogical support provided by learning software (3) and scientific research into learning and learners (4). Student modeling in general or, as stated by Baker and Yacef, improving student models is one of most cited domains of research in EDM. We have unfolded this group into more detailed categories in the previous section. Improving domain models based on the objective of application (which is our focus) can be known as part of decision support

Table 1 Targeted Users of EDM Applications

	Students	Educators	Administrators	Researchers
Predicting performance and characteristics		X	X	
Detecting undesirable student behaviour		X	X	
Profiling and Grouping students	X	X		
Social Network Analysis		X	X	X
Providing reports	X	X	X	
Creating alerts for stakeholders		X	X	
Planning and scheduling	X	X	X	
Constructing courseware		X		
Developing Concept Maps		X		X
Generating recommendation	X	X		
Evaluation		X		
Adaptive systems	X			
Scientific inquiry				X

systems such as developing concept maps or proving reports. Also, studying pedagogical support and scientific research can be summarized as scientific inquiry.

In another survey about EDM applications written by Romero and Ventura (2010) and published in 2010, 11 categories of application were suggested based on 300 research studies completed before 2010. This survey has been extremely useful as a reference for this paper, as it provides many examples for each of the introduced categories as well as methods and techniques used in them. The categories of applications introduced in this survey are:

- Analysis and Visualization of Data
- Providing Feedback for Supporting Instructors
- Recommendations for Students
- Predicting Student's Performance
- Student Modeling
- Detecting Undesirable Student Behaviors
- Grouping Students
- Social Network Analysis
- Developing Concept Maps
- Constructing Courseware
- Planning and Scheduling

These applications are all mentioned in our list of applications with a few changes and additions. A part of the end objective analysis and data visualization is to provide reports (or feedback, as mentioned above). Moreover, providing reports (or feedback) is not simply limited to supporting instructors; it can also target students and administrators.

In another survey (Romero and Ventura 2013) published in 2013 by the same authors, the same list, with a few changes, is introduced. The main topic presented in this survey is the idea of categorizing applications and tasks based on the end user. This is potentially helpful in developing a better understanding of the problems related to EDM. The end users proposed by Romero and Ventura are learners, educators, administrators and researchers. We have used the same classification in our report. We have also specified the possible targeted end users of each application based on the same classification.

The Handbook of Educational Data Mining (Romero et al. 2010) was published in 2010, and discusses some of the research done in the field of EDM. The focus of this book is not to categorize EDM applications, although it does mention a few of them, namely communicating to stakeholders, maintaining and improving courses, generating recommendation, predicting student grades and learning outcomes, student modeling and domain structure analysis. These applications have been discussed in more details in the previous works of the authors (Romero, Ventura, Pechenizkiy and Baker), as mentioned earlier.

In 2013, the book *Educational Data Mining: Applications and Trends* (Peña-Ayala 2013) was published. It includes an introduction to the field which discusses different applications, methods and datasets. In the second, third and fourth part of the book, there are some case studies categorized in three groups (student modeling, assessment and trends). The general goals of EDM are introduced in the first part of the book to show the wide range of applications. The following topics are discussed.

- Student modeling
- Predicting student performance and learning outcomes
- Generating recommendation
- Analyzing learner's behavior
- Communicating to stakeholders
- Domain structure analysis
- Maintaining and improving courses
- Studying the effects of pedagogical support that can be provided by learning software
- Advancing scientific knowledge about learning and learners through building, discovering or improving models of the student.

This list shares many topics with previous works, but the case studies in the third part of the book entitled "Assessment", add something new. These cover evaluation in cases in which evaluation is not an easy task. Improving evaluation as a possible application of EDM had not been addressed before. That was one important reason for us to specify a category of applications under "Evaluation", as discussed earlier.

Finally in one of the most recent surveys on EDM in higher education (Hegazi and Abugroon n.d.) published in 2016, four categories of applications have been counted, namely course management systems, student behaviors, decision support system in higher education and student retention and attrition. This list of applications introduces a number of EDM tasks, but it its treatment is not as clear, nor does it cover wide range of applications. For example, student retention can be considered as a part of student behavior and there are many different topics in course management systems which may result in different applications.

To conclude, we would like to compare the different categories of applications in various surveys and books with each other in the Fig. 2. In this figure, we show what applications each of the cited references cover. Some of the labels for categories of applications may vary, but in the end, they address the same applications using a different terminology. For example, we have equated domain structure analysis, cited in a few sources (Peña-Ayala 2013; Romero et al. 2010). with constructing the concept maps. The definitions provided for these concepts are not exactly the same; however, based on the examples provided by the references, we concluded that they represent the same group of applications. This figure shows the topics that have been mentioned more often than others or the topics for which there has been less focus as on others.

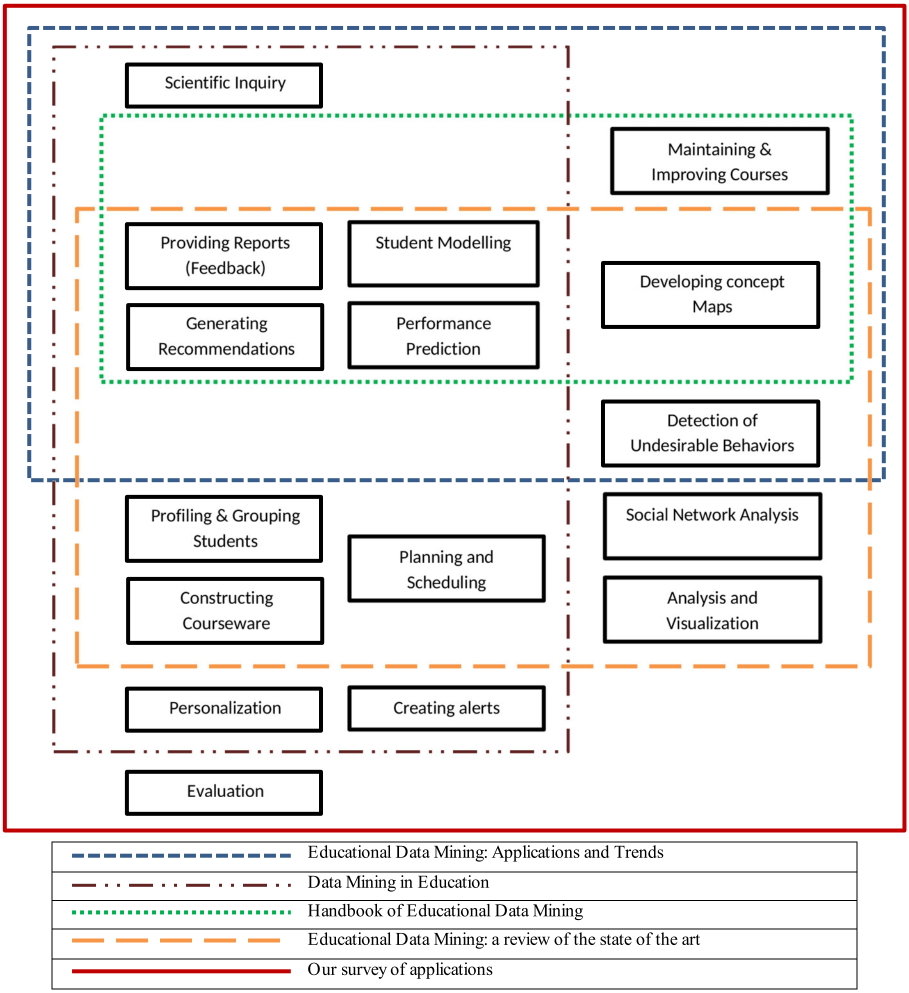
Scientific Inquiry
Maintaining & Improving Courses
Providing Reports (Feedback)
Student Modelling
Developing concept Maps
Generating Recommendations
Performance Prediction
Detection of Undesirable Behaviors
Profiling & Grouping Students
Social Network Analysis
Planning and Scheduling
Constructing Courseware
Analysis and Visualization
Personalization
Creating alerts
Evaluation
Educational Data Mining: Applications and Trends
Data Mining in Education
Handbook of Educational Data Mining
Educational Data Mining: a review of the state of the art
Our survey of applications

Fig. 2 Comparison of Reported applications in EDM

8 Conclusion

In this study, we reviewed the existing surveys and books about EDM and integrated the tasks introduced in each of them. We studied the examples provided in the surveys and books, as well as the publications of recent years in the *Journal of Educational Data Mining* to see if our proposed list covered all the research results. We grouped similar applications into categories and sub-categories in order to propose a taxonomy of tasks in EDM. This classification of applications/tasks is based on the end objectives. We chose some representative examples for each category. These examples can help us understand the categories better. However, they do not cover all the possible tasks in each category. We compared our proposed list of applications with the existing publications. Our list of applications is more exhaustive in terms of EDM topics compared to previous surveys and books, and proposes a novel and better suited categorization in a growing field. With the growth of computer based learning and availability of data, we believe the uses of EDM will also grow, leading to yet new applications.

References

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2014). Study navigator: An algorithmically generated aid for learning from electronic textbooks. *JEDM-Journal of Educational Data Mining*, *6*(1), 53–75.

Alaofi, M., & Rumantir, G. (2015). Personalisation of generic library search results using student enrolment information. *JEDM-Journal of Educational Data Mining*, *7*(3), 68–88.

Azarnoush, B., Bekki, J.M., Runger, G.C., Bernstein, B.L., & Atkinson, R.K. (2013). Toward a framework for learner segmentation. *JEDM-Journal of Educational Data Mining*, *5*(2), 102–126.

Baker, R.S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3–17.

Bates, A.W. (2015). Teaching in a digital age: Guidelines for designing teaching and learning. Tony Bates Associates.

Bravo, J., & Ortigosa, A. (2009). Detecting symptoms of low performance using production rules. In *International working group on educational data mining*.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729.

Cocea, M., & Weibelzahl, S. (2006). Can log files analysis estimate learners' level of motivation?

Dekker, G.W., Pechenizkiy, M., & Vleeshouwers, J.M. (2009). Predicting students drop out: A case study. In *International working group on educational data mining*.

Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M.R. (2008). Data mining application in higher learning institutions. *Informatics in Education-International Journal*, *7*, 31–54.

Galyardt, A., & Goldin, I. (2015). Move your lamp post: Recent data reflects learner knowledge better than older data. *JEDM-Journal of Educational Data Mining*, *7*(2), 83–108.

García, E., Romero, C., Ventura, S., & Castro, C. D. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction*, *19*(1–2), 99–132.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *JEDM-Journal of Educational Data Mining*, *7*(1), 33–50.

Harley, J.M., Trevors, G.J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *JEDM-Journal of Educational Data Mining*, *5*(1), 104–146.

Hegazi, M.O., & Abugroon, M.A. (n.d.) The state of the art on educational data mining in higher education.

Hsia, T.C., Shie, A.J., & Chen, L.C. (2008). Course planning of extension education to meet market demand by using data mining techniques—An example of Chinkuo technology university in Taiwan. *Expert Systems with Applications*, *34*(1), 596–602.

Huang, C.T., Lin, W.T., Wang, S.T., & Wang, W.S. (2009). Planning of educational training courses by data mining: Using China Motor Corporation as an example. *Expert Systems with Applications*, *36*(3), 7199–7209.

Kinnebrew, J.S., Loretz, K.M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM-Journal of Educational Data Mining*, *5*(1), 190–219.

Knowles, J.E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *JEDM-Journal of Educational Data Mining*, *7*(3), 18–67.

Lee, C.H., Lee, G.G., & Leu, Y. (2009). Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. *Expert Systems with Applications*, *36*(2), 1675–1684.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965.

Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). Defining ill-defined domains; a literature survey. In *Proceedings of the workshop on intelligent tutoring systems for ill-defined domains at the 8th international conference on intelligent tutoring systems* (pp. 1–10).

Macfadyen, L.P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599.

Mallavarapu, A., Lyons, L., Shelley, T., & Slattery, B. (2015). Developing computational methods to measure and track learners' spatial reasoning in an open-ended simulation. *JEDM-Journal of Educational Data Mining*, *7*(2), 49–82.

Miller, L.D., Soh, L.-K., Samal, A., Kupzyk, K., & Nugent, G. (2015). A comparison of educational statistics and data mining approaches to identify characteristics that impact online learning. *JEDM-Journal of Educational Data Mining*, *7*(3), 117–150.

Naveh, G., Tubin, D., & Pliskin, N. (2012). Student satisfaction with learning management systems: A lens of critical success factors. *Technology, Pedagogy and Education*, *21*(3), 337–350.

Novak, J.D., & Cañas, A.J. (2008). The theory underlying concept maps and how to construct and use them.

O'Mahony, M.P., & Smyth, B. (2007). A recommender system for on-line course enrolment: An initial study. In *Proceedings of the 2007 ACM conference on recommender systems* (pp. 133–136). ACM.

Peña-Ayala, A. (2013). *Educational data mining: Applications and trends* (Vol. 524). Berlin: Springer.

Rallo, R., Gisbert, M., & Salinas, J. (1999). Using data mining and social networks to analyze the structure and content of educative on-line communities. *Analysis*, *468*(472), 473.

Reffay, C., & Chanier, T. (2003). How social network analysis can help to measure cohesion in collaborative distance-learning. In (pp. 343–352). Springer.

Reyes, P., & Tchounikine, P. (2005). Mining learning groups' activities in Forum-type tools. In *Proceedings of th 2005 conference on computer support for collaborative learning: Learning 2005: The next 10 years!* (pp. 509–513). International Society of the Learning Sciences.

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. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *40*(6), 601–618.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S. (2010). *Handbook of educational data mining*. Boca Raton: CRC Press.

Romero, C., Zafra, A., Luna, J. M., & Ventura, S. (2013). Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Systems*, *30*(2), 162–172.

Sabourin, J.L., Rowe, J.P., Mott, B.W., & Lester, J.C. (2013). Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *JEDM-Journal of Educational Data Mining*, *5*(1), 9–38.

Self, J.A. (n.d.) Bypassing the intractable problem of student modelling.

Tang, C., Lau, R.W., Li, Q., Yin, H., Li, T., & Kilis, D. (2000). Personalized courseware construction based on web data mining. In *Proceedings of the first international conference on web information systems engineering, 2000* (Vol. 2, pp. 204–211). IEEE.

Vialardi, C., Agapito, J.B., Shafti, L.S., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In T. Barnes, M. Desmarais, C. Romero, S. Ventura.

Wang, V. C. (2014). *Handbook of research on education and technology in a changing society*. IGI Global.

Waters, A., Studer, C., & Baraniuk, R. (2014). Collaboration-type identification in educational datasets. *JEDM-Journal of Educational Data Mining*, *6*(1), 28–52.

Zimmermann, J., Brodersen, K.H., Heinimann, H.R., & Buhmann, J.M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *JEDM-Journal of Educational Data Mining*, *7*(3), 151–176.