Educational Data Mining: A Review of the State of the Art
Cristóbal Romero, Member, IEEE, and Sebastián Ventura, Senior Member, IEEE

Abstract—Educational data mining (EDM) is an emerging inter-disciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date. First, it introduces EDM and describes the different groups of user, types of educational environments, and the data they provide. Then it goes on to list the most typical/common tasks in the educational environment that have been resolved through data-mining techniques, and finally, some of the most promising future lines of research are discussed.

Index Terms—Data mining (DM), educational data mining (EDM), educational systems, knowledge discovery.

I. INTRODUCTION

EDUCATIONAL data mining (EDM) is a field that exploits statistical, machine-learning, and data-mining (DM) algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues [27]. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn [21]. On one hand, the increase in both instrumental educational software as well as state databases of student’s information have created large repositories of data reflecting how students learn [143]. On the other hand, the use of Internet in education has created a new context known as e-learning or web-based education in which large amounts of information about teaching–learning interaction are endlessly generated and ubiquitously available [60]. All this information provides a gold mine of educational data [186]. EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners. EDM has emerged as a research area in recent years for researchers all over the world from different and related research areas, which are as follows.

1) Offline education try to transmit knowledge and skills based on face-to-face contact and also study psychologically on how humans learn. Psychometrics and statistical techniques have been applied to data, like student’s behavior/performance, curriculum, etc., that was gathered in classroom environments.

2) E-learning and learning management system (LMS). E-learning provides online instruction, and LMS also provides communication, collaboration, administration, and reporting tools. Web mining (WM) techniques have been applied to student’s data stored by these systems in log files and databases.

3) Intelligent tutoring system (ITS) and adaptive educational hypermedia system (AEHS) are an alternative to the just-put-it-on-the-web approach, trying to adapt teaching to the needs of each particular student. DM has been applied to data picked up by these systems, such as log files, user models, etc.

The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. This process does not differ much from other application areas of DM, like business, genetics, medicine, etc., because it follows the same steps as the general DM process [219]: preprocessing, DM, and postprocessing. However, it is important to note that in this paper, the term DM is used in a larger sense than the original/traditional DM definition, i.e., we are going to describe not only EDM studies that use typical DM techniques, such as classification, clustering, association-rule mining, sequential mining, text mining, etc., but also describe other approaches, such as regression, correlation, visualization, etc., which are not considered to be DM in a strict sense.Furthermore, some methodological innovations and trends in EDM, such as discovery with models and the integration of psychometric modeling frameworks, are unusual DM categories or are not necessarily seen universally as being DM [20].

From a practical point of view, EDM allows, for example, to discover new knowledge based on students’ usage data in order to help to validate/evaluate educational systems, to potentially improve some aspects of the quality of education, and to lay the groundwork for a more effective learning process [219]. Some similar ideas were already successfully applied in e-commerce systems, the first and most popular application of DM [211], in order to determine clients’ interests so as to be able to increase online sales. However, to date, there has been comparatively less progress in this direction in education, although this situation is changing and there is currently an increasing interest in applying DM to the educational environment [228]. Even so, there are
some important issues that differentiate the application of DM, specifically to education, from how it is applied in other domains [221].

1) **Objective:** The objective of DM in each application area is different. For example, in EDM, there are both applied research objectives, such as improving the learning process and guiding students’ learning, as well as pure research objectives, such as achieving a deeper understanding of educational phenomena. These goals are sometimes difficult to quantify and require their own special set of measurement techniques.

2) **Data:** In educational environments, there are many different types of data available for mining. These data are specific to the educational area, and therefore have intrinsic semantic information, relationships with other data, and multiple levels of meaningful hierarchy. Some examples are the domain model, used in ITS and AEHS, which represents the relationships among the concepts of a specific subject in a graph or hierarchy format (e.g., a course consists of several chapters that are organized in lessons and each lesson includes several concepts); and the Q-matrix that shows relationships between items/questions of a test/quiz system and the concepts evaluated by the test. Furthermore, it is also necessary to take pedagogical aspects of the learner and the system into account.

3) **Techniques:** 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 cannot and have to be adapted to the specific educational problem at hand. Furthermore, specific DM techniques can be used for specific educational problems.

EDM involves different groups of users or participants. Different groups look at educational information from different angles, according to their own mission, vision, and objectives for using DM [104]. For example, knowledge discovered by EDM algorithms can be used not only to help teachers to manage their classes, understand their students’ learning processes, and reflect on their own teaching methods, but also to support a learner’s reflections on the situation and provide feedback to learners [177]. Although an initial consideration seems to involve only two main groups, the learners and the instructors, there are actually more groups involved with many more objectives, as can be seen in Table I.

Nowadays, there is a great variety of educational systems/environments such as: the traditional classroom, e-learning, LMS, adaptive hypermedia (AH) educational systems, ITS, tests/quizzes, texts/contents, and others such as: learning object (LO) repositories, concept maps, social networks, forums, educational game environments, virtual environments, ubiquitous computing environments, etc. All data provided by each of the aforementioned educational environments are different, thus enabling different problems and tasks to be resolved using DM techniques (see Section II). Table II shows a list of the most important studies on EDM grouped according to the type of data/environment involved.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>EDM USERS/STAKEHOLDERS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users/Actors</strong></td>
<td><strong>Objectives for using data mining</strong></td>
</tr>
<tr>
<td>Learners/Students/Pupils</td>
<td>To personalize e-learning; to recommend activities to learners’ resources and learning tasks that could further improve their learning; to suggest interesting learning experiences to the students; to suggest path pruning and shortening or simply links to follow, to generate adaptive hints, to recommend courses, relevant discussions, etc.</td>
</tr>
<tr>
<td>Educators/Teachers/Instructors/Tutors</td>
<td>To get objective feedback about instruction; to analyze students’ learning and behavior; to detect which students require support; to predict student performance; to classify learners into groups; to find a learner’s regular as well as irregular patterns; to find the most frequently made mistakes; to determine more effective activities; to improve the adaptation and customization of courses, etc.</td>
</tr>
<tr>
<td>Course Developers/Educational Researchers</td>
<td>To evaluate and maintain courseware; to improve student learning; to evaluate the structure of course content and its effectiveness in the learning process; to automatically construct student models and tutor models; to compare data mining techniques in order to be able to recommend the most useful one for each task; to develop specific data mining tools for educational purposes; etc.</td>
</tr>
<tr>
<td>Organizations/Learning Providers/Universities/Private Training Companies</td>
<td>To enhance the decision processes in higher learning institutions; to streamline efficiency in the decision-making process; to achieve specific objectives; to suggest certain courses that might be valuable for each class of learners; to find the most cost-effective way of improving retention and grades; to select the most qualified applicants for graduation; to help to admit students who will do well in university, etc.</td>
</tr>
<tr>
<td>Administrators/School District Administrators/Network Administrators/System Administrators</td>
<td>To develop the best way to organize institutional resources (human and material) and their educational offer; to utilize available resources more effectively; to enhance educational program offers and determine the effectiveness of the distance learning approach; to evaluate teacher and curricula; to set parameters for improving web-site efficiency and adapting it to users (optimal server size, network traffic distribution, etc.).</td>
</tr>
</tbody>
</table>

On the other hand, the International Working Group in EDM (http://www.educationaldatamining.org) has achieved the establishment of an annual International Conference on EDM in 2008, EDM’08 [19], EDM’09 [27], and EDM’10 [22]. This conference has evolved from previous EDM workshops at the AIED’07 [112], the EC-TEL’07 [222], the ICALT’07 [35], the UM’07 [17], the AAII’06 [34], the ITS’06 [111], the AAAT’05 [33], the AIED’05 [62], the ITS’04 [32], and the ITS’00 [30] conferences.

The number of publications about EDM has grown exponentially in the past few years (see Fig. 1). A clear sign of this tendency is the appearance of the peer-reviewed Journal of Educational Data Mining (JEDM) and two specific books on EDM edited by Romero and Ventura entitled: Data Mining in E-learning [220] and The Handbook of Educational Data Mining [228] co-edited by Baker and Pechenizkiy. There were also two surveys carried out previously about EDM. The first one [221] is a former review of Romero and Ventura with 81 references until 2005 in which papers were classified by the DM techniques used. In fact, this survey is an improved, updated, and much extended version of this previous one with 306 references.
TABLE II
LIST OF EDM REFERENCES GROUPED ACCORDING TO TYPES OF DATA USED

<table>
<thead>
<tr>
<th>Type of Data/Environment</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Education</td>
<td>[32], [42], [66], [68], [78], [94], [97], [102], [117], [118], [121], [128], [131], [139], [140], [145], [146], [162], [163], [167], [173], [195], [196], [210], [215], [236], [237], [239], [252], [258], [261], [269], [271], [278], [280], [304].</td>
</tr>
<tr>
<td>Web-based Education/E-learning</td>
<td>[11], [45], [49], [59], [63], [64], [85], [91], [96], [99], [101], [103], [116], [120], [127], [130], [144], [147], [151], [153], [154], [155], [156], [157], [175], [179], [180], [181], [188], [191], [197], [199], [212], [214], [225], [238], [240], [246], [253], [259], [263], [272], [275], [276], [284], [285], [286], [288], [289], [292], [293], [295], [298], [300].</td>
</tr>
<tr>
<td>Learning Management Systems</td>
<td>[28], [46], [48], [59], [67], [76], [100], [109], [110], [132], [159], [164], [168], [171], [178], [182], [183]. [208], [209], [223], [224], [232], [242], [254], [266], [267], [274], [291], [303].</td>
</tr>
<tr>
<td>Intelligent Tutoring Systems</td>
<td>[9], [15], [16], [18], [26], [29], [31], [47], [61], [65], [83], [98], [106], [114], [124], [134], [143], [174], [177], [185], [200], [203], [213], [217], [218], [234], [249], [265], [280], [287], [294].</td>
</tr>
<tr>
<td>Adaptive Educational Systems</td>
<td>[4], [23], [37], [38], [69], [92], [93], [105], [123], [125], [133], [136], [138], [148], [160], [161], [187], [219], [227], [245], [257], [260], [268], [277], [279], [301].</td>
</tr>
<tr>
<td>Tests/Questionnaires/Contents</td>
<td>[7], [12], [14], [25], [41], [43], [51], [54], [57], [79], [88], [126], [165], [194], [201], [202], [204], [205], [248], [270], [281], [283], [302].</td>
</tr>
<tr>
<td>Others</td>
<td>[1], [3], [40], [73], [107], [141], [150], [158], [235], [247], [251], [264], [283], [297].</td>
</tr>
<tr>
<td></td>
<td>[2], [13], [44], [53], [55], [71], [74], [77], [108], [122], [137], [142], [152], [190], [198], [206], [216], [231], [233], [250], [262], [299].</td>
</tr>
</tbody>
</table>

in which papers are classified by educational categories/tasks and the types of data used. It also shows some examples of new categories that have appeared since the 2005 survey, such as social network analysis and constructing courseware. The other survey [20] is a recent review by Baker and Yacef with 46 references encompassing up to 2009. This survey uses mainly the top eight most cited papers in the first 2005 review and the Proceedings of the EDM’08 and the EDM’09 conferences; it also groups papers according to EDM methods and applications, as we describe in Section II.

Finally, it is important to highlight that most of the pioneer and older research (from 1993 to 1999) deals with predicting student’s performance (see Task D in Section II). In fact, there is a huge body of studies on this topic in educational journals and conferences; and although seminal works date back to decades ago, new developments are highly relevant.

This survey is organized as follows. Section II lists the most common tasks in education that have been resolved by using DM techniques. Section III describes some of the most prominent future research lines. Finally, conclusions are outlined in Section IV.

II. EDUCATIONAL TASKS AND DM TECHNIQUES

There are many applications or tasks in educational environments that have been resolved through DM. For example, Baker [20], [21] suggests four key areas of application for EDM: improving student models, improving domain models, studying the pedagogical support provided by learning software, and scientific research into learning and learners; and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. Castro et al. [60] suggests the following EDM subjects/tasks: applications dealing with the assessment of the student’s learning performance, applications that provide course adaptation and learning recommendations based on the student’s learning behavior, approaches dealing with the evaluation of learning material and educational web-based courses, applications that involve feedback to both teacher and students in e-learning courses, and developments for detection of atypical students’ learning behaviors. However, as we think that there are even more possible applications, we have established our own categories (see Fig. 2) for the main educational tasks that have employed DM techniques. These categories come from

Fig. 1. Number of published papers until 2009 grouped according to the year. Note that we have counted only 300 papers in our reference section and not the total number of papers that were really published about EDM.

Fig. 2. Number of published papers until 2009 grouped by task/category. Note that we have counted only 300 papers in our reference section and not the total number of papers actually published about EDM.
different research communities (as we have previously described in Section I), and they also use different DM tasks and techniques. On one hand, we can see in Table II that the most active communities are e-learning/LMS and ITS/AEHS. On the other hand, we will see in the following sections that the most commonly applied DM tasks are regression, clustering, classification, and association-rule mining; and the most used DM techniques/methods are decision trees, neural networks, and Bayesian networks.

As we can see in Fig. 2, the categories or research lines that have the most papers published are the first eight ones (from A to G with 23 or more references each), and the categories that have the fewest papers published are the last four (from H to K with less than 15 references). We think that this may be mainly due to the fact that the first eight categories are older than the last four (and so more authors have worked on these tasks), but it could also be because of the special interest in each one. For example, although social network analysis is one of the newest tasks, it has more papers than the other three. We also want to point out that we have organized these categories by grouping them near the most closely related ones, which in our opinion are the following: since tasks A and B provide information to instructors and C to the students; D, E, F, and G tasks reveal students’ characteristics; H and I study graphs and relationships between students and concepts, respectively; and J and K help in creating/planning courseware and the course, respectively. Next, we are going to describe in detail these tasks/categories and the most relevant studies. But, as there are closely related areas, some references could be located in a different category or in several.

A. Analysis and Visualization of Data

The objective of the analysis and visualization of data is to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students’ course activities and usage information to get a general view of a student’s learning. Statistics and visualization information are the two main techniques that have been most widely used for this task.

Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data [86]. It is relatively easy to get basic descriptive statistics from statistical software, such as SPSS. Used with educational data, this descriptive analysis can provide such global data characteristics as summaries and reports about learner’s behavior [282]. It is not surprising that teachers prefer pedagogically oriented statistics (overall success rate, mastery levels, typical misconceptions, percentage of exercises tackled, and material read) that are easy to interpret [301]. On the other hand, teachers find the fine-grained statistics in log data too cumbersome to inspect or too time-consuming to interpret. Statistical analysis of educational data (logs files/databases) can tell us things such as: where students enter and exit, the most popular pages, the browsers students tend to use, and patterns of use over time, [130]; the number of visits, origin of visitors, number of hits, and patterns of use throughout various time periods [95]; number of visits and duration per quarter, top search terms, and number of downloads of e-learning resources [99]; number of different pages browsed and total time for browsing different pages [127]; usage summaries and reports on weekly and monthly user trends and activities [183]; session statistics and session patterns [199]; statistical indicators on the learner’s interactions in forums [5]; the amount of material students might go through and the order in which students study topics [212]; resources used by students and resources valued by students [241]; the overall averages of contributions to discussion forums, the amount of posting versus replies, and the amount of learner-to-learner interaction versus learner-to-teacher interaction [110]; the time a student dedicates to the course or a particular part of it [199]; the learners’ behavior and time distribution and the distribution of network traffic over time [303]; and the frequency of studying events, patterns of studying activity, timing and sequencing of events, and the content analysis of students’ notes and summaries [103]. Statistical analysis is also very useful to obtain reports assessing [81] how many minutes the student has worked, how many minutes he has worked today, how many problems he has resolved, and his correct percentage, our prediction of his score, and his performance level.

Information visualization uses graphic techniques to help people to understand and analyze data [172]. Visual representations and interaction techniques take advantage of the human eye’s broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once. There are several studies oriented toward visualizing different educational data such as: patterns of annual, seasonal, daily, and hourly user behavior on online forums [40]; the complete educational (assessment) process [205]; mean values of attributes analyzed in data to measure mathematical skills [302]; tutor–student interaction data from an automated reading tutor [185]; statistical graphs about assignments complement, questions admitted, exam score, etc. [242]; student tracking data regarding social, cognitive, and behavioral aspects of students [170]; student’s attendance, access to resources, overview of discussions, and results on assignments and quizzes [171]; weekly information regarding students’ and groups’ activity [135]; student’s progression per question as a transition between the types of questions [38]; fingertip actions in collaborative learning activities [11]; deficiencies in a student’s basic understanding of individual concepts [286] and higher education student-evaluation data [131]; student’s interactions with online learning environments [132]; the students’ online exercise work, including students’ interactions and answers, mistakes, teachers’ comments, etc. [176]; questions and suggestions in an adaptive tutorial [39]; navigational behavior and the performance of the learner [37]; educational trails of Web pages visited and activities done [225]; and the sequence of LOs and educational trails [238].

B. Providing Feedback for Supporting Instructors

The objective is to provide feedback to support course authors/teachers/administrators in decision making (about how to improve students’ learning, organize instructional resources
more efficiently, etc.) and enable them to take appropriate proactive and/or remedial action. It is important to point out that this task is different than data analyzing and visualizing tasks, which only provide basic information directly from data (reports, statistics, etc.). Moreover, providing feedback divulges completely new, hidden, and interesting information found in data. Several DM techniques have been used in this task, although association-rule mining has been the most common. Association-rule mining reveals interesting relationships among variables in large databases and presents them in the form of strong rules, according to the different degrees of interest they might present [296].

There are many studies that apply/compare several DM models that provide feedback. Association rules, clustering, classification, sequential pattern analysis, dependency modeling, and prediction have been used to enhance web-based learning environments to improve the degree to which the educator can evaluate the learning process [292]. Association analysis, clustering analysis, and case-based reasoning have also been used to organize course material and assign homework at different levels of difficulty [243]. Clustering, classification, and association-rule mining have been applied to develop a service to allow the evaluator to gather feedback from the learning progress automatically, and thus, appraise online course effectiveness [232]. Decision trees, Bayesian models, and other prediction techniques have been proposed to address the admission and counseling process in order to assist in improving the quality of education and student’s performance [215]. Several classifier algorithms have been applied to predict whether the teacher will recommend an intervention strategy for motivational profiles [124]. Clustering and association rules have been used in the academic community to potentially improve some qualitative teaching aspects [271].

Association-rule mining has been used to confront the problem of continuous feedback in the educational process [208]; to analyze learning data and to figure out whether students use resources and possibly whether their use has any (positive) impact on marks [178]; to determine the relationship between each learning-behavior pattern so that the teacher can promote collaborative learning behavior on the Web [289]; to find embedded information, which can be provided to teachers to further analyze, refine, or reorganize teaching materials and tests in adaptive learning environments [260]; to optimize the content of the university e-learning portal [214]; to discover interesting associations between student attributes, problem attributes, and solution strategies in order to improve online education systems for both teachers and students [181]; to analyze rule-evaluation measures in order to discover the most interesting rules [267]; to identify interesting and unexpected learning patterns, which in turn may provide decision lines, enabling teachers to more efficiently organize their teaching structure [272]; to provide feedback to the course author about how to improve courseware [219]; to analyze the user’s access log in Moodle to improve e-e-learning and to support the analysis of trends [28]; to find relationships between students’ LMS access behavior and overall performances in order to understand student’s web-usage patterns [46]; to improve an adaptive course design in order to show recommendations on how to enhance the course structure and contents [268]; to find interesting relationships between attributes, solution strategies adopted by learners, etc., from a web-based mobile learning system [299]; to help the teacher to discover beneficial or detrimental relationships between the use of web-based educational resources and student’s learning [226]; to reveal information about university students’ enrollment [236]; to help organizations to determine the thinking styles of learners and the effectiveness of a website structure [101]; to evaluate educational website design [164]; and to mine open answers in questionnaire data in order to analyze surveys [283].

Other different DM techniques have been applied to provide feedback such as: domain-specific interactive DM to find the relationships between log data and student’s behavior in an educational hypermedia system [123]; temporal DM to describe, interpret, and predict student’s behavior, and to evaluate progress in relation to learning outcomes in ITSs [29]; learning decomposition and logistic regression to compare the impact of different educational interventions on learning [84]; timely alerts to detect critical teaching and learning patterns and to help teachers to make sense of what is happening in their classrooms [246]; and usage data analysis to improve the effectiveness of the learning process in e-learning systems [182].

A special type of feedback is when data come specifically from tests, questions, assessments, etc. In this case, the objective is to analyze it in order to improve the questionnaires and to answer questions such as: what items/questions test the same information, and which are of the most use for predicting course/test results, etc. Several DM approaches and techniques (clustering, classification, and association analysis) have been proposed for joint use in the mining of student’s assessment data [204]. A group of DM techniques, i.e., statistic correlation analysis, fuzzy clustering analysis, grey relational analysis, K-means clustering, and fuzzy-association-rule mining have been applied to support mobile formative assessment in order to help teachers to understand the main factors influencing learner’s performance [55]. Several clustering algorithms (K-means, agglomerative clustering, and spectral clustering) have been applied to extract underlying relationships from a score matrix in order to help instructors to generate a large unit test [248]. Hierarchical clustering has been used for mining multiple-choice assessment data for similarity of the concepts represented by the responses [165]. Common-factor analysis and collaborative filtering have been used to discover the fundamental topics of a course from item-level grades [281]. Association-rule mining has been applied to analyze questionnaire data by discovering rule patterns in questionnaire data [54].

Finally, another special type of feedback involves the use of text data. In this case, the objective of applying text/DM to educational data is to analyze educational contents, to summarize/analyze the learner’s discussion process, etc., in order to provide instructor feedback. Automatic text analysis, content analysis, and text mining have been used to extract and identify the opinions found on Web pages in e-learning systems [247]; to mine free-form spoken responses given to tutor prompts by estimating the probability that a response has of mentioning a given target or set of targets [297]; to facilitate the automatic
coding process of an online discussion forum [158]; for collaborative learning prompted by learners’ comments on discussion boards [264]; to assess asynchronous discussion forums in order to evaluate the progress of a thread discussion [73]; and to identify patterns of interaction and their sequential organization in computer-supported collaborative environments like chats [44].

C. Recommendations for Students

The objective is to be able to make recommendations directly to the students with respect to their personalized activities, links to visits, the next task or problem to be done, etc., and also to be able to adapt learning contents, interfaces, and sequences to each particular student. Several DM techniques have been used for this task, but the most common are association-rule mining, clustering, and sequential pattern mining. Sequence/sequential pattern mining aims to discover the relationships between occurrences of sequential events to find if there exists any specific order in the occurrences [70].

Sequential pattern mining has been developed to personalize recommendations on learning content based on learning style and web-usage habits [298]; to study eye movements (of students’ reading concept maps) in order to detect when focal actions overlap unrelated actions [192]; for developing personalized learning scenarios in which the learners are assisted by the system based on patterns and preferred learning styles [23]; to identify significant sequences of activity indicative of problems/success in order to assist student teams by early recognition of problems [137]; to generate personalized activities for learners [277]; for personalizing based on itineraries and long-term navigational behavior [184]; to recommend the most appropriate future links for a student to visit in a web-based adaptive educational system [227]; to include the concept of recommended itinerary in Sharable Content Object Reference Model (SCORM) standard by combining teachers’ expertise with learned experience [184]; to select different LOs for different learners based on learner’s profiles and the internal relation of concepts [244]; for personalizing activity trees according to learning portfolios in a SCORM compliant environment [277]; for recommending lessons (LOs or concepts) that a student should study next while using an AH system [148]; to discover LO relationship patterns to recommend related LOs to learners [198]; and for adapting learning resource sequencing [136].

Association-rule mining has been used to recommend online learning activities or shortcuts on a course website [293]; to produce recommendations for learning material in e-learning systems [166]; for content recommendation based on educationally contextualized browsing events for web-based personalized learning [274]; for recommending relevant discussions to the students [2]; to provide students with personalized learning suggestions by analyzing their test results and test-related concepts [57]; for making recommendations to courseware authors about how to improve adaptive courses [92]; for building a personalized e-learning material-recommender system to help students to find learning materials [160]; for course recommendation with respect to optimal elective courses [253]; and for designing a material recommendation system based on the learning actions of previous learners [159].

Clustering has been developed to establish a recommendation model for students in similar situations in the future [276]; for grouping Web documents using clustering methods in order to personalize e-learning based on maximal frequent item sets [251]; for providing personalized course material recommendations based on learner’s ability [161]; and to recommend to students those resources they have not yet visited, but would find most helpful [96].

Other DM techniques used are: neural networks and decision trees to provide adaptive and personalized learning support [100]; production rules to help students to make decisions about their academic itineraries [269]; decision tree analysis to recommend optimal learning sequences to facilitate the students’ learning process and maximize their learning outcome [279]; learning factor transfers and Q-matrixes to generate domain models that will sequence item types to maximize learning [203]; an item-order effect model to suggest the most effective item sequences to facilitate learning [202]; a fuzzy item-response theory to recommend appropriate courseware for learners [50]; intelligent agent technology and SCORM-based course objects to build an agent-based recommender system for lesson plan sequencing in web-based learning [284]; DM and text mining to recommend books related to the books that the target pupil has consulted [189]; case-based reasoning to offer contextual help to learners, providing them with an adapted link structure for the course [114]; Markov decision process to automatically generate adaptive hints in ITS (to identify the action that will lead to the next state with the highest value) [249]; and an extended serial blog article composition particle swarm optimization (SBACPSO) algorithm to provide optimal recommended materials to users in blog-assisted learning [122].

D. Predicting Student’s Performance

The objective of prediction is to estimate the unknown value of a variable that describes the student. In education, the values normally predicted are performance, knowledge, score, or mark. This value can be numerical/continuous value (regression task) or categorical/discrete value (classification task). Regression analysis finds the relationship between a dependent variable and one or more independent variables [72]. Classification is a procedure in which individual items are placed into groups based on quantitative information regarding one or more characteristics inherent in the items and based on a training set of previously labeled items [75]. Prediction of a student’s performance is one of the oldest and most popular applications of DM in education, and different techniques and models have been applied (neural networks, Bayesian networks, rule-based systems, regression, and correlation analysis).

A comparison of machine-learning methods has been carried out to predict success in a course (either passed or failed) in ITSs [106]. Other comparisons of different DM algorithms are made to classify students (predict final marks) based on Moodle usage data [224]; to predict student’s performance (final grade) based...
on features extracted from logged data [180]; and to predict university students’ academic performance [128].

Different types of neural-network models have been used to predict final student grades (using back-propagation and feedforward neural networks) [94]; to predict the number of errors a student will make (using feedforward and backpropagation) [280]; to predict performance from test scores (using backpropagation and counter propagation) [78]; to predict students’ marks (pass or fail) from Moodle logs (using radial basis functions) [67]; and for predicting the likely performance of a candidate being considered for admission into the university (using multilayer perceptron topology) [196].

Bayesian networks have been used to predict student—applicant performance [102]; to model user knowledge and predict student’s performance within a tutoring system [200]; to predict a future graduate’s cumulative grade point average based on applicant background at the time of admission [117]; to model two different approaches to determine the probability a multiskill question has of being corrected [201] and to predict future group performance in face-to-face collaborative learning [250]; to predict end-of-year exam performance through student’s activity with online tutors [12]; and to predict item response outcome [69].

Different types of rule-based systems have been applied to predict student’s performance (mark prediction) in an e-learning environment (using fuzzy-association rules) [191]; to predict learner’s performance based on the learning portfolios compiled (using key-formative assessment rules) [51]; for prediction, monitoring, and evaluation of student’s academic performance (using rule induction) [195]; to predict final grades based on features extracted from logged data in an education web-based system (using genetic algorithm to find association rules) [240]; to predict student’s grades in LMSs (using grammar-guided genetic programming) [291]; to predict student’s performance and provide timely lessons in web-based e-learning systems (using decision tree) [45]; and to predict online students’ marks (using an orthogonal search-based rule extraction algorithm) [76].

Several regression techniques have been used to predict students’ marks in an open university (using model trees, neural networks, linear regression, locally weighed linear regression, and support vector machines) [146]; for predicting end-of-year accountability assessment scores (using linear regression prediction models) [7]; to predict student’s performance from log and test scores in web-based instruction (using a multivariable regression model) [288]; for predicting student’s academic performance (using stepwise linear regression) [97]; for predicting time to be spent on a learning page (using multiple linear regression) [8]; for identifying variables that could predict success in colleges courses (using multiple regression) [167]; for predicting university students’ satisfaction (using regression and decision trees analysis) [258]; for predicting exam results in distance education courses (using linear regression) [188]; for predicting when a student will get a question correct and association rules to guide a search process to find transfer models to predict a student’s success (using logistic regression) [88]; to predict the probability a student has of giving the correct answer to a problem in an ITS (using a robust ridge regression algorithm) [61]; for predicting end-of-year accountability assessment scores (using linear regression) [7]; to predict a student’s test score (using stepwise regression) [79]; and to predict the probability that the student’s next response has of being correct (using linear regression) [31].

Finally, correlation analyses have been applied together to predict web-student performance in online classes [275]; to predict a student’s final exam score in online tutoring [207]; and for predicting high school students’ probabilities of success in university [173].

E. Student Modeling

The objective of student modeling is to develop cognitive models of human users/students, including a modeling of their skills and declarative knowledge. DM has been applied to automatically consider user characteristics (motivation, satisfaction, learning styles, affective status, etc.) and learning behavior in order to automate the construction of student models [89]. Different DM techniques and algorithms have been used for this task (mainly, Bayesian networks).

Several DM algorithms (naïve Bayes, Bayes net, support vector machines, logistic regression, and decision trees) have been compared to detect student mental models in ITSs [234]. Unsupervised (clustering) and supervised (classification) machine learning have been proposed to reduce development costs in building user models and to facilitate transferability in intelligent learning environments [4]. Clustering and classification of learning variables have been used to measure the online learner’s motivation [115].

Bayesian networks have been used to make predictions about student’s knowledge, i.e., the probability that student has of knowing a skill at a given time through cognitive tutors [18]; to detect students’ learning styles in a web-based education system [91]; to predict whether a student will answer a problem correctly [134]; to model a student’s changing state of knowledge during skill acquisition in ITS [47]; to infer unobservable learning variables from students’ help-seeking behavior in a web-based tutoring system [10]; and for knowledge tracing in order to verify the impact of self-discipline on students’ knowledge and learning [98].

Sequential pattern mining has been used to automatically acquire the knowledge to construct student models [9]; to identify meaningful user characteristics and to update the user model to reflect newly gained knowledge [6]; and for predicting students’ intermediate mental steps in sequences of actions stored by—learning environments based on problem solving [218].

Association-rule algorithms have been applied for personality mining based on web-based education models in order to deduce learners’ personality characteristics [120] and for student modeling in ITSs [168].

Other DM techniques and models have also been used for student modeling. A logistic regression model has been used to construct transfer models (to accurately predict the level at which a student represents knowledge) [83]. A learning agent that models student behaviors using linear regression has been constructed in order to predict the probability that the student’s
next response has of being correct [31]. Inductive logic programming and a profile extractor system (using numeric algorithms) have been developed to induce student profiles in e-learning systems [155]. The Markov decision process has been proposed to automatically create student models by generating hints for an IT that learns [26]. Fuzzy techniques have used student models in web-based learning environments in order to generate advice for the teachers [144]. A dynamic learning-response model has been developed for inferring, testing, and verifying student’s learning models on an adaptive learning website [125]. Bootstrapping novice data can create an initial skeletal model of a tutor from log data collected from actual use of the tool by students [174]. A collaborative-based DM approach has been developed for diagnostic and predictive student modeling purposes in integrated learning environments [151]. Multiple correspondence analysis and cross validation by correlation analysis have been applied to identify learning styles in Index of Learning Styles (ILS) questionnaires [270]. The Q-matrix method has been used to create concept models that represent relationships between concepts and questions, and to group student’s test question responses according to concepts [25]. An algorithm to estimate Dirichlet priors has been developed to produce model parameters that provide a more plausible picture of student’s knowledge [213]. Self-organizing maps and principal component analysis have been applied for predictive and compositional modeling of the student’s profile [150]. A clustering algorithm (K-means) has been developed to model student’s behavior with a very small set of parameters without compromising the behavior of the system [217].

F. Detecting Undesirable Student Behaviors

The objective of detecting undesirable student behavior is to discover/detect those students who have some type of problem or unusual behavior such as: erroneous actions, low motivation, playing games, misuse, cheating, dropping out, academic failure, etc. Several DM techniques (mainly, classification, and clustering) have been used to reveal these types of students in order to provide them with appropriate help in plenty of time.

Several of the classification algorithms that have been used to detect problematic student’s behavior are decision tree neural networks, naïve Bayes, instance-based learning, logistic regression, and support vector machines for predicting/preventing student drop out [145]; feed-forward neural networks, support vector machines, and a probabilistic ensemble simplified fuzzy ARTMAP algorithm to predict dropouts in e-learning courses [156]; Bayesian nets, logistic regression, simple logic classification, instance-based classification, attribute-selected classification, bagging, classification via regression, and decision trees for engagement prediction [64]; decision tree, Bayesian classifiers, logistic models, the rule-based learner, and random forest to detect/predict first-year student drop out [66]; paired t-test for grouping students by common misconceptions (hint-driven learners and failure-driven learners) [287]; C4.5 decision tree algorithm for detecting any potential symptoms of low performance in e-learning courses [41]; decision trees to identify students with little motivation [63]; decision trees for detection of irregularities and deviations in the learners’ actions in an interactive learning environment [187]; and the J48 decision tree algorithm and farthest-first clustering algorithm for predicting, understanding, and preventing academic failure (exam failure) among university students [42].

Different types of clustering also used to carry out this task are: Kohonen nets to detect students that cheat in online assessments [43]; outlier detection to uncover atypical student behavior [265]; an outlier detection method using Bayesian predictive distribution to detect learners’ irregular learning [263]; a constrained mixture of student r-distribution and generative topographic mapping to detect atypical student behavior (outliers) [59]; and an augmented version of the Levenshtein distance algorithm to identify novice errors and error paths [265].

Finally, other DM techniques and models used for this task are, for example, association-rule mining for selecting weak students for remedial classes [163], to send warning messages to students with unusual learning behavior in an AEHS [133], and to construct concept-effect relationships for diagnosing student’s learning problems [126]; a latent response model to identify if students are playing with the system (to detect student misuse) in a way that would lead to poor learning [15] and to automatically detect when a student is off-task in a cognitive tutor [16]; Bayesian networks to predict the need for help in an interactive learning environment [169]; stepwise regression to detect misplay and look for sources of error in the prediction of student’s test scores [79]; human reliability analysis to infer the underlying causes that lead to the production of trainee errors in a virtual environment [74]; and Markov chain analysis to identify and classify common student errors and technical problems in order to prevent them from occurring in the future [109].

G. Grouping Students

The objective is to create groups of students according to their customized features, personal characteristics, etc. Then, the clusters/groups of students obtained can be used by the instructor/developer to build a personalized learning system, to promote effective group learning, to provide adaptive contents, etc. The DM techniques used in this task are classification (supervised learning) and clustering (unsupervised learning). Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster have some points in common [229].

Different clustering algorithms have been used to group students such as: hierarchical agglomerative clustering, K-means and model-based clustering to identify groups of students with similar skill profiles [14]; a clustering algorithm based on large generalized sequences to find groups of students with similar learning characteristics based on their traversal path patterns and the content of each page they have visited [256]; model-based clustering to automatically discover useful groups from LMS data to obtain profiles of student’s behavior [254]; a hierarchical clustering algorithm for user modeling (learning styles) in intelligent e-learning systems in order to group students according to their individual learning style preferences [294]; discriminating features and external profiling features (pass/fail) to
support teachers in collaborative student modeling [90]; an improvement in the matrix-based clustering method for grouping learners by characteristics in e-learning [295]; a fuzzy clustering algorithm to find interested groups of learners according to their personality and learning strategy data collected from an online course [259]; a hybrid method of clustering and Bayesian networks to group students according to their skills [105]; a K-means clustering algorithm for effectively grouping students who demonstrate similar learning portfolios (students’ assignment scores, exam scores, and online learning records) [51]; an expectation–maximization algorithm to form heterogeneous groups according to student’s skills [188]; a K-means clustering algorithm to discover interesting patterns that characterize the work of stronger and weaker students [209]; a conditional subspace clustering algorithm to identify skills that differentiate students [194]; a two-step cluster analysis to classify how students organize personal information spaces (piling, one-folder, small-folders, and big-folder filing) [108]; hierarchical cluster analysis to establish the proportion of students who get an exercise wrong or right [24]; and a genetic clustering algorithm to solve the problem of allocating new students (which places new students into classes so that the gaps between learning levels in each class is minimum and the number of students in each class does not exceed the limit) [304].

Several classification algorithms have been applied in order to group students such as: discriminant analysis, neural networks, random forests, and decision trees for classifying university students into three groups (low-risk, medium-risk, and high-risk of failing) [252]; classification and regression tree, chi-squared automatic interaction detection, and C4.5 algorithm for the automatic identification of the students’ cognitive styles [153]; a classification and regression tree to create a decision tree model to illustrate a user’s learning behavior, in order to analyze it according to different cognitive style groups [151]; a hidden-Markov-model-based classification approach to characterize different types of users through their navigation or content access patterns [85]; decision trees for classifying students according to their accumulated knowledge in e-learning systems [179]; C4.5 decision tree algorithm for discovering potential student groups with similar characteristics who will react to a particular strategy [49]; naïve Bayes classifier to classify learning styles that describe learning behavior and educational content [138]; genetic algorithms for grouping students according to their profiles in a peer review content [65]; classification trees and multivariate adaptive regression to identify those students who tend to take online courses and those who do not [290]; decision tree and support vector machine for assessing an activity by more than one lecturer using a pairwise learning model [210]; a classification algorithm for speech act patterns to assess participants’ roles and identify discussion threads [141]; and K-nearest neighbor (K-NN) classification combined with genetic algorithms to identify and classify student learning styles [48].

I. Developing Concept Maps

The objective of constructing concept maps is to help instructors/educators in the automatic process of developing/constructing concept maps. A concept map is a conceptual graph that shows relationships between concepts and expresses the hierarchical structure of knowledge [193]. Some DM techniques (mainly, association rules, and text mining) have been used to construct concept maps.
Association-rule mining has been used to automatically construct concept maps guided by learners’ historical testing records [262]; to discover concept-effect relationships for diagnosing the learning problems of students [126]; and for conceptual diagnosis of e-learning through automatically constructed concept maps that enable teachers to overcome the learning barrier and misconceptions of learners [152].

Text mining has been applied to automatically construct concept maps from academic articles in the e-learning domain [52]; to formulate concept maps from online discussion boards using fuzzy ontology [149]; to find relationships between text documents and construct document index graphs [107]; and to explore cognitive concept-map differences in instructional outcomes [119].

Finally, a specific concept-map algorithm has been created to automatically organize knowledge points and map them [243]; a method of automatic concept relationship discovery for an adaptive e-course has been developed to help teachers to author overall automation [245]; and a multiexpert e-training course design model has been developed by concept-map generation in order to help the experts to organize their domain knowledge [58].

J. Constructing Courseware

The objective of constructing courseware is to help instructors and developers to carry out the construction/development process of courseware and learning contents automatically. On the other hand, it also tries to promote the reuse/exchange of existing learning resources among different users and systems.

Different DM techniques and models have been used to develop courseware. The clustering of students and naïve algorithms have been proposed to construct personalized courseware by building a personalized Web tutor tree [255]. Rough set theory and clustering concept hierarchy have been used to construct e-learning frequently asked questions (FAQ) retrieval infrastructures [56], Multilingual knowledge-discovery technique processing has been combined with AH techniques to automatically create online information systems from linear texts in electronic format, such as textbooks [3]. Argument mining has been proposed to support argument construction for agents and ITSs using different mining techniques [1].

Several DM techniques have been applied to reuse learning resources. Hybrid unsupervised DM techniques have been employed to facilitate LO reuse and retrieval from the Web or from different LO repositories [142]. Valuable information can be found by mining metadata from educational resources (ontology of pedagogical objects), which helps DM to retrieve more precise information for content reuse and exchange [175]. The automatic classification of Web documents in a hierarchy of concepts based on naïve Bayes has been suggested for the indexing and reuse of learning resources [235]. Profile analysis based on collaborative filtering has been used to search LOs and rank search results according to the predicted level of user interest [197]. Mining educational multimedia presentations has been used to establish explicit relationships among the data related to interactivity (links and actions) and to help to predict interactive properties in the multimedia presentations [13].

K. Planning and Scheduling

The objective of planning and scheduling is to enhance the traditional educational process by planning future courses, helping with student course scheduling, planning resource allocation, helping in the admission and counseling processes, developing curriculum, etc. Different DM techniques have been used for this task (mainly, association rules).

Classification, categorization, estimation, and visualization have been compared in higher education for different objectives, such as academic planning, predicting alumni pledges, and creating meaningful learning outcome typologies [162]. Decision trees, link analysis, and decision forests have been used in course planning to analyze enrollees’ course preferences and course completion rates in extension education courses [118]. Classification, prediction, association-rule analysis, clustering, etc., have been compared to discover new explicit knowledge that could be useful in the decision-making process in higher learning institutions [68]. Educational training courses have been planned through the use of cluster analysis, decision trees, and back-propagation neural networks in order to find the correlation between the course classifications of educational training [121]. Decision trees and Bayesian models have been proposed to help management institutes to explore the probable effects of changes in recruitments, admissions, and courses [215].

Association-rule mining has been used to provide new, important, and therefore, demand-oriented impulses for the development of new bachelor and master courses [237]. Curriculum revision has been done by association-rule mining in order to identify and understand whether curriculum revisions can affect students in a university [36]. A decisional tool (based on association-rule mining) has been constructed to help in making decisions on how to improve the quality of the service provided by the university based on students’ success and failure rates [239]. Association-rule mining and genetic algorithms have been applied to an automatic course-scheduling system to produce the course timetables that best suit student and teacher needs [278].

Finally, a regression model has been developed to predict the likelihood a specific undergraduate applicant has of matriculation, if admitted [139]; several clustering algorithms (self-organizing map networks, K-means, and kth-nearest neighbor) have been used as a decision support in selecting Association of Advance Collegiate Schools of Business (AACSB) peer schools [140].

III. FUTURE WORK AND RESEARCH LINES

Although there is a lot of future work to be considered in EDM, we indicate in continuation what arguably are the most interesting and influential among them. In fact, a few initial studies on some of these points have already begun to appear.

1) EDM tools have to be designed to be easier for educators or nonexpert users in DM. DM tools are normally designed more for power and flexibility than for simplicity. Most
of the current DM tools are too complex for educators to use and their features go well beyond the scope of what an educator may want to do. For example, on one hand, users have to select the specific DM method/algorithm they want to apply/use from the wide range of methods/algorithms available on DM. On the other hand, most of the DM algorithms need to be configured before they are executed. Users have to provide appropriate values for the parameters in advance in order to obtain good results/models, and therefore, the user must possess a certain amount of expertise in order to find the right settings. One possible solution is the development of wizard tools that use a default algorithm for each task and parameter-free DM algorithms to simplify the configuration and execution for nonexpert users. EDM tools must also have a more intuitive interface that is easy to use and with good visualization facilities to make their results meaningful to educators and e-learning designers [93]. It is also very important to develop specific preprocessing tools in order to automate and facilitate all the preprocessing functions or tasks that EDM users currently must do manually.

2) Integration with the e-learning system. The DM tool has to be integrated into the e-learning environment as one more traditional authoring tool (course creator, test creator, report tools, etc.). All DM tasks (preprocessing, DM, and postprocessing) must be carried out in a single application with a similar interface. In this way, EDM tools will be more widely used by educators, and feedback and results obtained with DM techniques could be easily and directly applied to the e-learning environment using an iterative evaluation process [224].

3) Standardization of data and models. Current tools for mining data pertaining to a specific course/framework may be useful to their developers only. There are no general tools or reusing tools that can be applied to any educational system. Therefore, a standardization of input data and output model are needed, as along with preprocessing, discovering, and postprocessing tasks. Shen et al. [243] proposed using Extensible Markup Language (XML) as data specification. Ventura et al. [267] used Predictive Modeling Markup Language (PMML) that is the leading standard for statistical and DM models. But, it is also necessary to incorporate domain knowledge and semantics using ontology-specification languages, such as Ontology Web Language (OWL) and Resource Description Framework (RDF); and standard metadata for e-learning, such as SCORM. In this line, currently, there is only one public educational data repository, the PSLC DataShop [143], which provides a lot of educational datasets and also facilitates analysis. However, all this log data are obtained from ITSs; therefore, it is necessary to have more public datasets from other types of educational environments as well. In this way, specific educational benchmark datasets could be used to compare/evaluate different DM algorithms.

4) Traditional mining algorithms need to be tuned to take into account the educational context. DM techniques must use semantic information when applied to educational data. This shows the need for more effective mining tools that integrate educational domain knowledge into DM algorithms. For example, Iksal and Choquet [129] have proposed specific usage tracking language (UTL) to describe the track semantics recorded by an LMS and to link them to the need for observation defined in a predictive scenario. Education-specific mining techniques can greatly improve instructional design and pedagogical decisions, and the aim of the semantic Web is to facilitate data management in educational environments.

IV. Conclusion

This paper is a review of the state of the art with respect to EDM and surveys the most relevant work in this area to date. In fact, after first collecting and consulting all the published bibliography in EDM area, we have selected each author’s most important studies. Then, we have classified each study not only by the type of data and DM techniques used, but also and more importantly, by the type of educational task that they resolve.

EDM has been introduced as an upcoming research area related to several well-established areas of research, including e-learning, AH, ITSs, WM, DM, etc. We have seen how fast EDM is growing as reflected in the increasing number of contributions published every year in international conferences and journals and the number of specific tools specially developed for applying DM algorithms in educational data/environments. Therefore, it could be said that EDM is now approaching its adolescence, i.e., it is no longer in its early days, but is not yet a mature area. In fact, we have described some interesting future lines, but for it to become a more mature area, it is also necessary for researchers to develop more unified and collaborative studies instead of the current plethora of multiple individual proposals and lines. Thus, the full integration of DM in the educational environment will become a reality, and fully operative implementations (both commercial and free) could be made available not only for researchers and developers, but also for external users.

REFERENCES


Cristóbal Romero (M’09) received the B.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1996 and 2003, respectively. He is currently an Associate Professor in the Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain. He has authored or coauthored more than 30 papers about educational data mining in international journals and conferences. His research interests include applying data mining in e-learning systems.

Dr. Romero is a member of the IEEE Computer Society, the International Educational Data Mining (EDM) Working Group, and the steering committee of the EDM Conferences.

Sebastián Ventura (SM’07) received the B.Sc. and Ph.D. degrees in sciences from the University of Córdoba, Córdoba, Spain, in 1989 and 1996, respectively. He is currently an Associate Professor in the Department of Computer Science and Numerical Analysis, University of Córdoba, where he is the Head of the Knowledge Discovery and Intelligent Systems Research Laboratory. He has authored or coauthored more than 70 international publications, 20 of them published in international journals. He has also worked on eleven research projects (being the Coordinator of two of them) that were supported by the Spanish and Andalusian governments and the European Union. His research interests include soft computing, machine learning, data mining, and its applications.

Dr. Ventura is a Senior Member of the IEEE Computer Society, the IEEE Computational Intelligence Society, the IEEE Systems, Man, and Cybernetics Society, and the Association of Computing Machinery.