Expert Systems with Applications 41 (2014) 1432–1462

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Educational data mining: A survey and a data mining-based analysis of recent works

Alejandro Peña-Ayala *

WOLNM & ESIME Zacatenco, Instituto Politécnico Nacional, U. Profesional Adolfo López Mateos, Edificio Z-4, 2do piso, cubiculo 6, Miguel Othón de Mendizábal S/N, La Escalera, Gustavo A. Madero, D.F., C.P. 07320, Mexico

ARTICLE INFO

Keywords:
Data mining
Educational data mining
Data mining profile
Educational data mining approach pattern
Pattern for descriptive and predictive educational data mining approaches

ABSTRACT

This review pursues a twofold goal, the first is to preserve and enhance the chronicles of recent educational data mining (EDM) advances development; the second is to organize, analyze, and discuss the content of the review based on the outcomes produced by a data mining (DM) approach. Thus, as result of the selection and analysis of 240 EDM works, an EDM work profile was compiled to describe 222 EDM approaches and 18 tools. A profile of the EDM works was organized as a raw data base, which was transformed into an ad-hoc data base suitable to be mined. As result of the execution of statistical and clustering processes, a set of educational functionalities was found, a realistic pattern of EDM approaches was discovered, and two patterns of value-instances to depict EDM approaches based on descriptive and predictive models were identified. One key finding is: most of the EDM approaches are ground on a basic set composed by three kinds of educational systems, disciplines, tasks, methods, and algorithms each. The review concludes with a snapshot of the surveyed EDM works, and provides an analysis of the EDM strengths, weakness, opportunities, and threats, whose factors represent, in a sense, future work to be fulfilled.

1. Introduction

Data mining (DM[1]) is a computer-based information system (CBIS) (Vlahos, Ferratt, & Knoepfle, 2004) devoted to scan huge data repositories, generate information, and discover knowledge. The meaning of the traditional mining term biases the DM grounds. But, instead of searching natural minerals, the target is *knowledge*. DM pursues to find out data patterns, organize information of hidden relationships, structure association rules, estimate unknown items' values to classify objects, compose clusters of homogenous objects, and unveil many kinds of findings that are not easily produced by a classic CBIS. Thereby, DM outcomes represent a valuable support for decisions-making.

Concerning education, it is a novel DM application target for knowledge discovery, decisions-making, and recommendation (Vialardi-Sacin, Bravo-Agapito, Shafti, & Ortigosa, 2009). Nowadays, the use of DM in the education arena is incipient and gives birth to the *educational data mining* (EDM) research field (Anjewierden, Kollöffel, & Hulshof, 2007). As we will see in Section 2, in a sense the first decade of the present century represents the kick-off of EDM.

EDM emerges as a paradigm oriented to design models, tasks, methods, and algorithms for exploring data from educational settings. EDM pursues to find out patterns and make predictions that characterize learners' behaviors and achievements, domain knowledge content, assessments, educational functionalities, and applications (Luan, 2002). Source information is stored in repositories managed by conventional, open, and distance educational modalities.

Some of the EDM trends are anticipated here. One of them corresponds to the standard integration of an EDM module to the typical architecture of the wide diversity of computer-based educational systems (CBES). Other tendency demands that EDM provides several functionalities during three stages of the teaching-learning cycle. The first stage corresponds to the provision of EDM *proactive* support for adapting the educational setting according to the student's profile prior to deliver a lecture. During

* Address: WOLNM & ESIME Zacatenco, Instituto Politécnico Nacional, U. Profesional Adolfo López Mateos, Edificio Z-4, 2do piso, cubiculo 6, Miguel Othón de Mendizábal S/N, La Escalera, Gustavo A. Madero, D.F., C.P. 07320, Mexico. Tel.: +52 55 5694 0916/+52 55 5454 2611 (cellular); fax: +52 55 5694 0916.
E-mail address: apenaa@ipn.mx
URL: http://www.wolnm.org/apa

[1] AIWBES: adaptive and intelligent web-based educational systems; BKT: Bayesian knowledge tracing; CBES: computer-based educational systems; CBIS: computer-based information system,; DM: data mining; DP: dynamic programming; EDM: educational data mining; EM: expectation maximization; HMM: hidden Markov model; IBL: instances-based learning; IRT: item response theory; ITS: intelligent tutoring systems; KDD: knowledge discovery in databases; KT: knowledge tracing; LMS: learning management systems; SNA: social network analysis; SWOT: strengths, weakness, opportunities, and threats; WBC: web-based courses; WBES: web-based educational systems.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.042

the student-system interaction stage, it is desirable that EDM acquires log-data and interprets their meaning in order to suggest recommendations, which can be used by the CBES for personalizing services to users at real-time. In the next stage, EDM should carry out the evaluation of the provided education concerning: delivered services, achieved outcomes, degree of user's satisfaction, and usefulness of the resources employed. What is more, several challenges (i.e., targets, environments, modalities, functionalities, kinds of data, ...) wait to be tackled or have been recently considered by EDM, such as: big data, cloud computing, social networks, web mining, text mining, virtual 3-D environments, spatial mining, semantic mining, collaborative learning, learning companions, ...

The present work extends the period described by earlier surveys, summarized in Section 2.2, that cover from 1995 up to 2009. The aim is to preserve and update the chronicles of recent EDM development. The scope of the work is limited and provides a partial image of the EDM activity published in all celebrated events and available media. In spite of this, the work provides a snapshot of the EDM labor that several members have been achieving. Inclusively, it applies the essential subject, DM, to organize, analyze, and discuss the content of the overview. Such a policy is a novelty: to preach through example.

As result of the application of such a policy, the next four contributions are offered to be used by the EDM community: a DM profile, an EDM approach profile, a pattern for EDM approaches based on descriptive models, and a pattern for EDM approaches based on predictive models. The first facilitates the description of the DM baseline that supports an EDM approach. The second is useful to define the nature and baseline of an EDM approach. The third and four are patterns to design EDM approaches, which are useful as a reference to develop similar versions of descriptive and predictive models.

In this paper a survey of EDM works fulfilled from 2010 up to 2013 1st Qtr. is presented. In addition, the method followed for producing the overview is outlined in Section 2, as well as the gathered material is stated. A *sample* of 240 EDM works is summarized in Section 3. Such a collection is organized according to typical functionalities fulfilled by CBES that were found from the material. In Section 4, an analysis of the sampled works is provided to shape the recent status and evolution of the EDM field, and some EDM approach patterns are highlighted. Finally, the conclusions Section tailors a snapshot of the sample and a critical analysis of the EDM arena that are useful to inspire future work.

2. Method and materials

In this section, the method and the materials of the overview are described. The method is a framework devoted to gather and mine EDM works. The materials tailor the survey domain through five subjects: a reference to prior EDM reviews, the scope of the collected EDM works, a profile of DM, a summary of CBES, and the data representation of EDM approaches used for mining.

As a result of the method application, a *sample* of 240 *EDM works* published between 2010 and the first quarter of 2013 was gathered. It is made up of two *sub-samples*, one of 222 *EDM approaches* and another of 18 *EDM tools* (i.e., the first represents EDM applications and the second software). The sample symbolizes a valuable source that is used to provide a highlight of the EDM arena in Section 3 and a brief analysis in Section 4. Moreover, the sample is examined to produce statistics and discover some findings, which are illustrated in the following subsections as well as in Sections 3 to 5.

2.1. Framework applied for knowledge discovery of educational data mining works

The method used to carry out this survey is a framework designed to gather, analyze, and mine EDM works. It follows a workflow to lead the activities oriented to knowledge discovery in databases (KDD). The workflow is split into three stages. The development of each stage is achieved by three tasks. Thus, nine tasks compose the whole KDD workflow pictured in Fig. 1, whose purpose and outcomes are explained as follows:

The "EDM work collection" stage performs three tasks. The first task seeks source references that publish EDM works. As a result, a collection of EDM works is gathered. The second evaluates EDM works and produces an *EDM approach profile* per each chosen EDM work. The third analyzes the EDM approach profiles and organizes a *raw EDM database*.

The "data processing stage" encompasses the tasks labeled as fourth, fifth, and sixth in Fig. 1. The fourth task transforms the raw EDM database into an *ad-hoc EDM database* to facilitate statis-

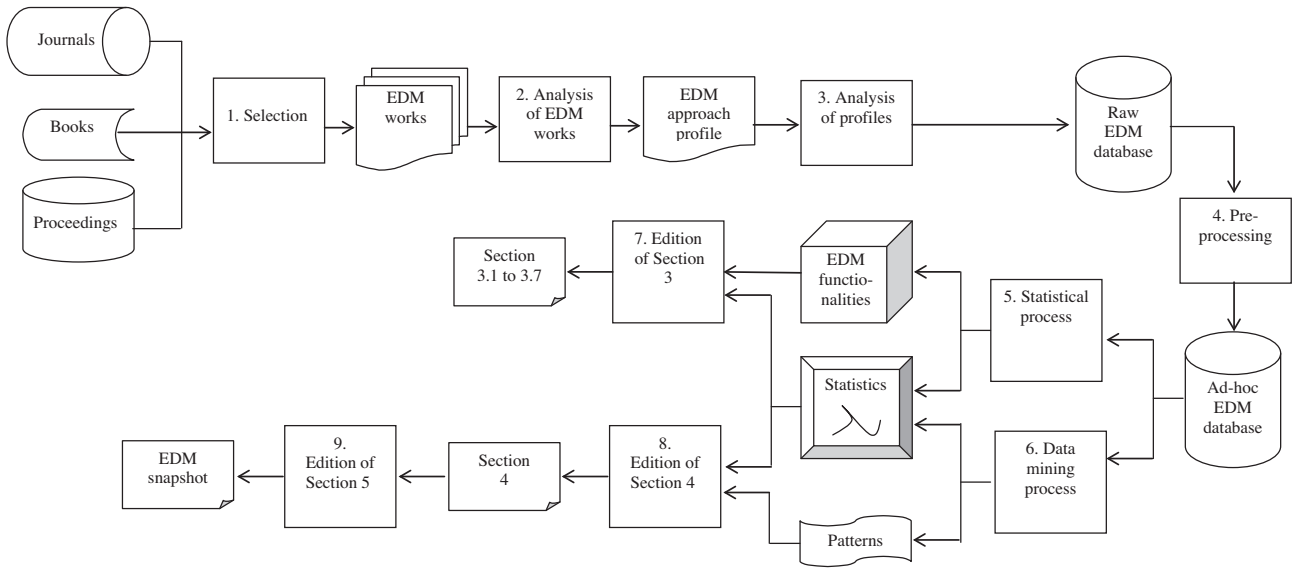

Fig. 1. Workflow of the DM approach performed to analyse, classify, represent, and mine data of the EDM related works.

tical and mining processes. The fifth performs statistical processes to generate seven kinds of EDM functionalities to gather homogeneous related works and statistics. The sixth mines the ad-hoc EDM database to find out patterns that characterize the gathered EDM works.

The stage oriented to "edit and interpret the results" contains the tasks labeled as seven to nine in Fig. 1. The seventh task classifies the EDM works according to the educational functionalities they most focus on. As result, seven topics are organized in balanced proportions of homogeneous EDM works to outline seven Sections presented as 3.1 to 3.7. The eighth interprets the patterns produced by the DM approach to discover relationships between the traits value-instances that characterize the EDM approaches. The last task analyzes the discovered knowledge from the EDM works to tailor a snapshot of the EDM arena that is described in Sections 2 to 5.

2.2. Previous reviews of data mining and educational data mining

As the starting point of this work, prior reviews of DM and EDM were examined to tailor a conceptual frame about the domain study. Therefore, five reviews are introduced in this subsection, where one is oriented to DM, and the other four cover the period from 1995 up to 2009.[2]

As EDM is based on DM, a review of DM techniques and applications achieved during 2000 to 2011 is summarized as follows. Shu-Hsien, Pei-Hui, and Pei-Yuan (2012) present a state of the art about DM that concerns a series of works fulfilled throughout the past decade. The paper surveys and classifies 216 works using nine categories that are presented with their respective counting of works: (a) neural networks: 9; (b) algorithm architecture: 22, (c) dynamic prediction: 17; (d) analysis of system architecture: 23; (e) intelligent agent systems: 14; (f) modeling: 15; (g) knowledge-based systems: 19; (h) systems optimization: 14; (i) information systems: 28. The authors recognize the broad baseline that supports DM models, tasks, methods, techniques, and algorithms. Finally, three suggestions are made: (1) include social sciences methodologies; (2) integrate several methodologies into a holistic one; (3) change the policy to guide future development of DM.

Regarding EDM, Romero and Ventura (2007) present a review of 81 works published from 1995 up to 2005, where only seven correspond to the 1990s. They identify statistics-visualization and web mining as a couple of DM techniques to classify the application of DM to CBES. As for statistics, several tools are identified and seven EDM works cited. Concerning visualization, four works are referenced and one tool is recognized. Regarding web mining, it is split into three kinds of tasks: (1) clustering, classification, and outlier detection; (2) association rules and sequential pattern; (3) text mining. A sample of EDM works is given for each kind of task. However, the sample is partitioned into three variants of e-Learning systems: particular web-based courses (WBC), well-known learning management systems (LMS), and adaptive and intelligent web-based educational systems (AIWBES). So, nine collections of EDM works are provided in the review with the next statistic: (a) 15 works of clustering, classification, and outlier detection tasks split into: 3 WBC, 3 LMS, 9 AIWBES; (b) 14 papers about association rules and sequential pattern tasks divided into: 6 WBC, 4 LMS, 4 AIWBES; (c) 7 works related to text mining partitioned into: 4 WBC, 2 LMS, 1 AIWBES. As future trends, they demand: friendly EDM tools for non-technical users, the standardization of DM methods and data; the integration of DM functionalities in CBES, and the design of techniques devoted to EDM.

A couple of reviews appeared in 2009 to shape a state of the EDM. The first is the work made by Baker and Yacef (2009). They celebrate the nascent EDM research community, define DM and EDM, and provide 45 EDM references, where one corresponds to 1973, another to 1995, and one more to 1999. The review identifies some EDM targets, such as: student models, models of domain knowledge, pedagogical support, and impacts on learning; where 8, 4, 3, and 4 related works are respectively cited. The second review published in 2009 was presented by Peña-Ayala, Domínguez, and Medel (2009). It offers 91 references about three topics: CBES, DM, and EDM. Concerning the former, approaches such as computer-assisted instruction, intelligent tutoring systems (ITS), LMS, and web-based educational systems (WBES) are considered. Regarding DM, several models, tasks, and techniques are identified; where mathematical, rules-based, and soft computing techniques are the target of analysis. As for EDM works, they are organized into four functionalities: student modeling, tutoring, content, and assessment.

The fourth EDM review corresponds to Romero and Ventura (2010), who enhanced their prior EDM survey adding 225 works, keeping the former seven references of the 1990s, and including three papers published in 2010. One novelty concerns a list of 235 works classified and counted in the following way: 36 in traditional education, 54 WBES, 29 LMS, 31 ITS, 26 adaptive educational systems, 23 test-questionnaires, 14 text-contents, and 22 others. Concerning EDM applications, they are gathered into eleven educational categories with the next counting: (a) analysis and visualization of data: 35; (b) providing feedback for supporting instruction: 40; (c) recommendations for students: 37; (d) predicting students' performance: 76; (e) student modeling: 28; (f) detecting undesirable student behaviors: 23; (g) grouping students: 26; (h) social network analysis: 15; (i) developing concept maps: 10; (j) constructing courseware: 9; (k) planning and scheduling: 11. At the end of the review, authors assert: "EDM is now approaching its adolescence..."

2.3. Scope of the present survey of educational data mining works

An interpretation of the four EDM reviews published up to 2010 shows that: the current century represents the start of EDM, because near to 98% of the cited works have appeared since 2000. In consequence, EDM is living its teenage period. During its growth, EDM has shifted from isolated papers published in conferences and journals, to dedicated workshops,[3] an international conference on educational data mining,[4] a specialized journal of EDM,[5] a handbook (Romero, Ventura, Pechenizkiy, & Ryan, 2011), and a society of experts and partisans,[6] as well as one edited book (Romero & Ventura, 2006) and another in press (Peña-Ayala, 2013). This synergy reveals the increasing interest in EDM and is the main reason to update the review by means of the present survey.

Therefore, the scope of the current overview is constrained to a sample of representative EDM works published in journals,[7] chapter books related to EDM, as well as papers presented in EDM conferences and workshops. The chosen references have been published during the period from 2010 to the first quarter of 2013. In this way, the EDM chronicles are extended and refreshed.

[2] As none of the works that are cited by the five reviews is included in the references of this paper, readers are encouraged to seek such papers to analyze the EDM background.

[3] See https://pslcdatashop.web.cmu.edu/KDD2011/

[4] See http://edm2013.iismemphis.org/

[5] See http://www.educationaldatamining.org/JEDM/

[6] See http://www.educationaldatamining.org/

[7] Most of the journals are indexed by © Thompson Reuters Journal Citation Reports and are published by prestigious editorials such as: Elsevier, Springer, IEEE, and Sage.

Table 1
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according nine disciplines (i.e., due several approaches apply more than one discipline, the total counting is 271).

Disciplines	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Probability	1	101	37.27	101	37.27
2. Machine learning	1	90	33.21	191	70.48
3. Statistic	1	47	17.34	238	87.82
4. Dynamic programming	1	18	6.64	256	94.46
5. Others with counting from 5 to 9	1	5	1.85	261	96.31
6. Others with counting from 2 to 4	4	10	3.69	271	100.00
7. Others with counting 1	0	0	0.00	271	100.00
Total	9	271	100.00		

2.4. Data mining in a nutshell

According to Witten and Frank (2000), DM is the process oriented to extract useful and comprehensible knowledge, previously unknown, from huge and heterogeneous data repositories. Thus, the design of a DM work demands the instantiation of several characteristics to shape the approach, such as: disciplines to tailor the theoretical baseline, the sort of model to be built, tasks to perform, methods and techniques to mechanize the proposal, as well as the algorithms, equations, and frames (e.g., data structures, frameworks) to deploy the approach on computers and internet settings. In consequence, this subsection is oriented to define those attributes, provide some of their instances, and reveal the statistics of their occurrence among the *sub-sample* of 222 EDM approaches.

2.4.1. Disciplines involved in data mining

The DM baseline is grounded by disciplines such as: probability (Karegar, Isazadeh, Fartash, Saderi, & Navin, 2008), machine learning (Witten, Frank, & Hall, 2011), statistic (Hill & Lewicki, 2006), soft computing (Mitra & Acharya, 2003), artificial intelligence (Bhattacharyya & Hazarika, 2006), and natural language (McCarthy & Boonthum-Denecke, 2011). Concerning the sub-sample, Table 1 asserts: probability, machine learning, and statistic offer the grounds of 88% EDM approaches!

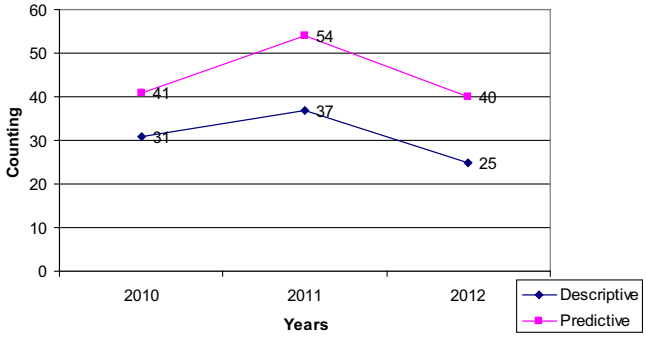

Fig. 2. Counting histogram of EDM approaches introduced in Sections 3.1 to 3.6, which are classified according two DM models (i.e., due several works apply more than one model the total counting is 231).

2.4.2. Data mining models

Essentially, two kinds of DM models are designed: descriptive and predictive (Kantardzic, 2011). Descriptive models usually apply unsupervised learning functions to produce patterns that explain or generalize the intrinsic structure, relations, and interconnectedness of the mined data (Peng, Kou, Shi, & Chen, 2008). Predictive models frequently apply supervised learning functions to estimate unknown or future values of dependent variables based on the features of related independent variables (Hand, Mannila, & Smyth, 2001). As for the sub-sample, Fig. 2 shows a three years tendency, from 2010 up to 2012, where 60% of the approaches depicts predictive models and 40% shapes descriptive models.

2.4.3. Data mining tasks

Usually, the implementation of a model is made by a task. For instance, clustering (Berkhin, 2006), association rules (Hong, Lin, & Wang, 2003), correlation analysis (Hardoon, Shawe-Taylor, & Szedmak, 2004), produce descriptive models; whilst, classification (Chau, Cheng, Kao, & Ng, 2006), regression (Wu & Li, 2007), and categorization generate predictive models (Genkin, Lewis, & Madigan, 2007). As for the sub-sample, Table 2 informs that: the most typical tasks are classification and clustering because together they reach 69% of the DM tasks used by EDM approaches!

2.4.4. Data mining methods and techniques

Once the DM model and tasks have been defined, the methods and techniques to build the approach are chosen according to the discipline. For instance, Bayes theorem (Pardos & Heffernan, 2010b), decision trees (McCuaig & Baldwin, 2012), instances-based learning (IBL) (Brighton & Mellish, 2002), and hidden Markov model (HMM) (Lee & Brunskill, 2012) are the most popular methods used by the approaches of the sub-sample, as Table 3 shows. Whereas, logistic (D'Mello & Graesser, 2010), linear regression (González-Brenes & Mostow, 2010), frequencies (Merceron, 2011),

Table 2
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 10 tasks (i.e., due several approaches apply more than one task, the total counting is 242).

Tasks	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Classification	1	102	42.15	102	42.15
2. Clustering	1	65	26.86	167	69.01
3. Regression	1	37	15.29	204	84.30
4. Association rules	1	16	6.61	220	90.91
5. Others with counting from 5 to 11	2	17	7.02	237	97.93
6. Others with counting from 2 to 4	1	2	0.83	239	98.76
7. Others with counting 1	3	3	1.24	242	100.00
Total	10	242	100.00		

and hierarchical clustering (Huei-Tse, 2011) techniques provide support for 45% of the sub-sample, as Table 4 shows.

2.4.5. Algorithms, equations, and frames used for data mining

After some method and/or technique are chosen to solve a specific DM task, an algorithm, equation, and/or frame are implemented to mine the source data (Wu et al., 2008). According to the EDM approaches that compose the sub-sample, the most popular algorithms, equations, and frames are respectively shown in Tables 5–7. Where, they respectively unveil: K-means (Bian, 2010), expectation maximization (EM) (Pardos & Heffernan, 2010a), J48 (Baker et al., 2012), and Naive–Bayes (Anaya & Boticario, 2011a) are the top-four most deployed algorithms; statistical equations, including descriptive, (Baker & Gowda, 2010) are the most used equations; several versions of Bayesian networks (Xu & Mostow, 2011a) are the most popular frames.

2.5. A glance at educational systems

CBES have evolved during more than sixty years, as an object of research, experimentation, development, and application, as well as commercial purposes. They represent an alternative to *conventional* education systems that deploy academic programs *on-site*. CBES represent an attempt to automate instruction and follow

Table 3
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according a 52 methods (i.e., due many approaches do not apply any method and others use more than one, the total counting is 244).

Methods	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Bayes theorem	1	48	19.67	48	19.67
2. Decision trees	1	44	18.03	92	37.70
3. Instances-based learning	1	22	9.02	114	46.72
4. Hidden Markov model	1	20	8.20	134	54.92
5. Others with counting from 5 to 13	5	54	22.13	188	77.05
6. Others with counting from 2 to 4	11	25	10.25	213	87.30
7. Others with counting 1	32	31	12.70	244	100.00
Total	52	244	100.00		

Table 4
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 43 techniques (i.e., due many approaches do not apply any technique and others use more than one, the total counting is 112).

Techniques	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Logistic regression	1	20	17.86	20	17.86
2. Linear regression	1	13	11.61	33	29.46
3. Frequencies	1	10	8.93	43	38.39
4. Hierarchical clustering	1	7	6.25	50	44.64
5. Others with counting from 5 to 6	2	10	8.93	60	53.57
6. Others with counting from 2 to 4	10	25	22.32	85	75.89
7. Others with counting 1	27	27	24.11	112	100.00
Total	43	112	100.00		

Table 5
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 143 algorithms (i.e., due many approaches do not apply any algorithm and others use more than one, the total counting is 274).

Algorithms	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. K-means	1	19	6.93	19	6.93
2. Expectation maximization	1	15	5.47	34	12.41
3. J48	1	15	5.47	49	17.88
4. NaiveBayes	1	13	4.74	62	22.63
5. Others with counting from 5 to 9	8	51	18.61	113	41.24
6. Others with counting from 2 to 4	22	55	20.07	168	61.31
7. Others with counting 1	109	106	38.69	274	100.00
Total	143	274	100.00		

Table 6
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 40 equations (i.e., due many approaches do not apply any equation and others use more than one, the total counting is 78).

Equations	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Statistical	1	21	26.92	21	26.92
1. Descriptive statistical	1	6	7.69	27	34.62
2. Logistic regression	1	5	6.41	32	41.03
3. Others with counting 3	2	6	7.69	38	48.72
4. Others with counting 2	5	10	12.82	48	61.54
5. Others with counting 1	30	30	38.46	78	100.00
Total	40	78	100.00		

Table 7
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 18 frames (i.e., due many approaches do not apply any frame and others use more than one, the total counting is 40).

Frames	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Bayesian networks	1	16	40.00	16	40.00
2. Dynamical Bayesian networks	1	5	12.50	21	52.50
3. Others with counting 2	3	6	15.00	27	67.50
4. Others with counting 1	13	13	32.50	40	100.00
Total	18	40	100.00		

the transformation trends of computer and communication sciences. In order to shape a profile of CBES, a set of different kinds of CBES and their basic functionalities are given in this section. Both subjects are illustrated by the statistics estimated for the sub-sample of 222 EDM approaches introduced in Section 3.

2.5.1. Variety of computer-based education systems

During the evolution of CBES, many study domains, different application purposes, and styles of user-system interaction have emerged. The diversity of CBES focus on specific targets (e.g., instruction, learning, problem-solving, skills development, management of courseware, gaming), apply particular learning theories (e.g., objectivism, constructivism, socialist), deploy specific functionalities (e.g., individualized, personalized, workgroup, collaboration, adaptive, intelligent), use different technological platforms (e.g., mainframes, personal computers, internet, mobile, ubiquitous), and follow pedagogical practices (e.g., student-centered, situated, long-life, immersive, blended).

In consequence, specific terms have been coined to label an educational, pedagogical, and technological paradigm of CBES, such as: ITS that behave like problem-solving monitors, coaches, laboratory instruments, and consultants (Psotka, Massey, & Mutter, 1989); LMS that virtually support the routine of teachers in the classroom dedicated to publish course material, design examples to analyze and solve, and define auto-grade quizzes; AIWBES that pursue to intelligently adapt the curricula, content, sequencing, assessment, and support given to learners according their background, skills, and progress to meet their educational goals (Rebak, Blackmon, & Humphreys, 2000).

As regards the sub-sample of EDM approaches, Table 8 identifies 37 kinds of CBES, where ITS and LMS are the most prominent with 49% of the sample. Whereas, Table 9 shows specific instances for the earlier kinds of CBES such as: Algebra, ASSISTments, Algebra-Bridge, and Moodle; where the first three are instances of ITS, and the last is a case of LMS. These four instances are the most popular in the EDM arena and support 48% of the sub-sample.

2.5.2. Diversity of functionalities provided by computer-based educational systems

Other consequence of the evolution of CBES is the variety of *functionalities* provided to the users, such as: content of the domain knowledge, student modeling, assessment, sequencing of lectures, student assistance, teacher support, collaboration, and several more.

Concerning the sub-sample of EDM approaches, Table 10 highlights that: nearly 82% focuses on three versions of student modeling (e.g., behavior, performance, and general) and assessment; but, the complement, 18%, corresponds to two sets of functionalities (e.g., the first concerns to student support and feedback; the second embraces curriculum, domain knowledge, sequencing, and teacher support).

In addition, Fig. 3 shapes a histogram of the counting estimated for the functionalities of EDM approaches during three years. It reveals: similar tendencies for student modeling, student behavior modeling, and assessment; an increasing tendency for student performance modeling; opposed patterns for student support and feedback versus curriculum-domain knowledge-sequencing-

Table 8
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 37 educational systems.

Educational systems	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Intelligent tutoring system	1	88	39.64	88	39.64
2. Learning management system	1	20	9.01	108	48.65
3. Conventional education	1	20	9.01	128	57.66
4. Computer-based educational system	1	15	6.76	143	64.41
5. Others with counting from 5 to 9	5	31	13.96	174	78.38
6. Others with counting from 2 to 4	13	33	14.86	207	93.24
7. Others with counting 1	15	15	6.76	222	100.00
Total	37	222	100.00		

Table 9
Counting of EDM approaches introduced in Sections 3.1 to 3.6, which are organized according 37 specific instances of educational systems (i.e., due several approaches do not identify the name of the educational system, the total counting is 130).

Specific educational systems	Items	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Algebra	1	20	15.38	20	15.38
2. ASSISTments	1	19	14.62	39	30.00
3. Moodle	1	13	10.00	52	40.00
4. Algebra-Bridge	1	10	7.69	62	47.69
5. Others with counting from 5 to 9	1	5	3.85	67	51.54
6. Others with counting from 2 to 4	13	30	23.08	97	74.62
7. Others with counting 1	33	33	25.38	130	100.00
Total	37	130	100.00		

Table 10
Counting of EDM approaches organized according to six functionalities that are presented in Sections 3.1 to 3.6.

Functionalities that represent the EDM approaches introduced in Sections 3.1 to 3.6	Counting	Percentage (%)	Accumulative counting	Accumulative percentage (%)
1. Student behavior modeling	48	21.62	48	21.62
2. Student performance modeling	46	20.72	94	42.34
3. Assessment	45	20.27	139	62.61
4. Student modeling	43	19.37	182	81.98
5. Student support and feedback	21	9.46	203	91.44
6. Curriculum, domain knowledge, sequencing, teacher support	19	8.56	222	100.00
Total	222	100.00		

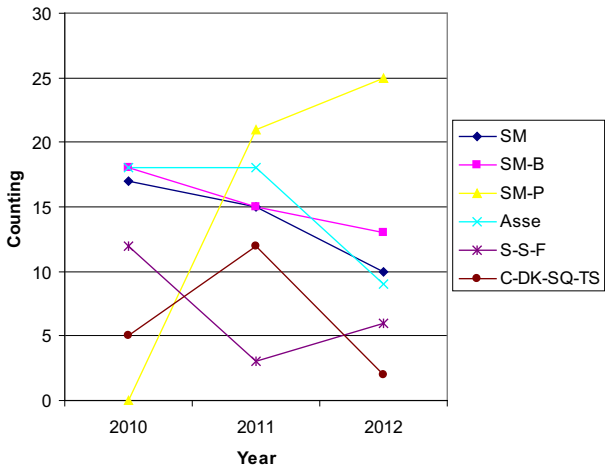

Fig. 3. Counting histogram of EDM approaches classified according to the six functionality subjects that label the Sections 3.1 to 3.6, whose counting is given in Table 10.

teaching support. They are respectively labeled in Fig. 3 as: SM, SM-B, Asse, SM-P, S-S-F, C-DK-SQ-TS.

Based on these statistics, Section 3 is organized into seven parts to present 240 *EDM works*, where Sections 3.1 to 3.6 summarize 222 *EDM approaches* and Section 3.7 depicts 18 *EDM tools*.

2.6. Data representation of educational data mining works

As result of applying the method for knowledge discovery of EDM works, three repositories are organized. The first is a set of *EDM works* that highlights the educational and DM traits stated in the printed version of chosen works. The second is a *raw EDM database* stored in the computer to represent the EDM approach profile, as well as a set of catalogs to standardize the educational and DM traits value-instances. The third is an *ad-hoc EDM database* that represents binary vectors to characterize the different value-instances for each educational and DM trait of the selected works.

As for the EDM work, it provides the raw source to define an *EDM approach profile*. Such a profile constitutes a basic record to depict an EDM approach. It embraces two kinds of traits to depict educational and DM characteristics of the approach. The first holds seven traits: functionality, role, role-type, module, module-type, system, and system-name. The second is a *DM profile* made up of eight traits: discipline, model, task, method, technique, algorithm, equation, and frame. The DM profile is used to characterize the sub-sample of 222 EDM approaches, where 150 are mature and 72 are incipient (i.e., due to they were still in *progress* at the year of their publication).

Concerning the raw EDM database, it represents the EDM approach profile through a spread-sheet with 15 columns and 2 more to identify the work and its publication year. Moreover, a set of catalogs provides a code to instantiate the values that depict the educational traits and the DM profile. As result, 15 catalogs are organized with the following example of instances: (1) *functionality*: 6 items to label the functionalities of the CBES (e.g., student modeling, assessment, student support); (2) *role*: 9 items identify the purpose of the EDM approach (e.g., assessment, domain knowledge, sequencing); (3) *role-type*: 40 instances specific targets of application (e.g., affect, behavior, cognition); (4) *module*: 8 values identify the CBES component (e.g., content, evaluation, tutoring); (5) *module-type*: 5 instances lead on the purpose of the module (e.g., advising, cognition, monitoring); (6) *system*: 37 options express the kind of educational system (e.g., WBES, ITS, LMS); (7) *system-name*: 51 terms label the educational system (e.g., ASSISTments, Andes, Moodle); (8) *discipline*: 9 items (e.g., machine learning, probability, statistic); (9) *model*: 2 options: descriptive and predictive; (10) *task*: 10 variants (e.g., association rules, classification, clustering); (11) *method*: 52 alternatives (e.g., Bayes theorem, cluster analysis, decision trees); (12) *technique*: 43 items (e.g., covariance, factorization, heuristics); (13) *algorithm*: 143 options (e.g., J48, Apriori, k-means); (14) *equation*: 40 options (e.g., linear algebra, mean squared error, time series); (15) *frame*: 18 instances (e.g., Akaike information criterion, Bayesian knowledge base, rough set model).

As result of pre-processing the raw EDM database, the integrity and consistency of the information stored in the ad-hoc EDM database is assured. Moreover, the EDM approach profile is transformed into a vector composed of alphanumerical items that label the traits values according to the code stated by their respective catalog. For instance, the items of vector [185, 2010, SM-B, SM, B, SMM, CO, ITS, Algebra, ML, D, CU, DT, k-means] respectively represent: paper-id, year of publication, student behavior modeling functionality, student modeling role, behavior role-type, student model module, cognition module-type, ITS system, Algebra system-name, machine learning discipline, descriptive model, clustering task, decision trees method, k-means algorithm. However, in order to facilitate the statistical and mining processes, the alphanumerical vector is converted into a binary row (e.g., [185, 2010, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,0,...]), where the first 1 identifies the student behavior modeling functionality "SM-B", and the second 1 depicts the student modeling role "SM", and so on.

3. Results

The development of the second stage that makes up the method for knowledge discovery produced a *sample* of 240 works that meet the criteria given in Section 2.3. In a sense, the sample reflects the EDM youth in progress. During the analysis of the sample it was found that most of the 222 approaches correspond to student modeling. In consequence, an imbalance in the former structure of the survey was noticed. So, in order to tailor a relevant organization of subjects that shows balanced sets of homogeneous works, called

functionalities and *tools*, a clustering process was achieved as follows:

Once the ad-hoc EDM database was generated, several descriptive statistics were estimated to produce a set of representative EDM functionalities. The main results are outlined in Table 11. It provides a snapshot of the works being presented in the following subsections. It is aggregated at three levels, the first represents the EDM works gathered in this survey, the second corresponds to the EDM approaches and EDM tools, and the third is reserved for their respective six functionalities and three kinds of tools. The first column identifies the aggregation level. The remaining columns reveal the number of published works per total, in progress, and mature, which are segmented by the period (e.g., 2010 to the first quarter of 2013) and the years (e.g., 2010, 2011,2012, and the first quarter of 2013).

As a result of the counting of several traits held by the EDM approach profile, seven balanced clusters of EDM works were organized. Such clusters are shown in Table 11; where six clusters represent functionalities of EDM approaches and one cluster gathers three kinds of EDM tools. The clusters that correspond to student modeling, student behavior modeling, student performance modeling, and assessment hold an average of 44.5 approaches (i.e., see the second column of Table 11); whereas the clusters devoted to student support and feedback, curriculum-domain knowledge-sequencing-teaching support, and tools reach an average of 19.3 approaches. The last cluster embraces several sub-clusters oriented to group different types of EDM tools (e.g., extraction-learning-feature, visualization, and analysis support), whose average is six works.

Unfortunately, due to the size of the sample and space constraints, the overview is tailored as a survey. Thus, six subsections are organized to present homogenous approaches according to six educational functionalities, and one subsection more is oriented to introduce EDM tools. Thereby, the subsections dedicated to describe functionalities are made up of an introduction and a summary. The introduction provides a definition of the functionality and identifies the number of the approaches that compose the survey. The summary points out a collection of related approaches. Moreover, with the aim at enhancing the summary, an analysis is presented in Section 4.1, where seven subsections (e.g., 4.1.1 to 4.1.7) briefly shape the state and evolution of the functionalities and tools highlighted in Sections 3.1 to 3.7.

Concerning the collection, it is split into four parts to present the approaches published in each year (i.e., from 2010 to the first quarter of 2013). So, each part contains the next subjects: (1) the number of the approaches published in a specific year; (2) a series of the approaches in progress; (3) a table to show the main traits of the EDM approach profile that characterizes mature approaches; (4) a series of mature approaches that are linked to the respective table by an *id*. Both series, of incipient and mature approaches, are highlighted with a brief profile of the approach due to the huge space that would be necessary to provide a full description or evaluation.

3.1. Student modeling

Student modeling is oriented to shape different domains that characterize the learner, such as: emotions, cognition, domain knowledge, learning strategies, achievements, features, learning preferences, skills, evaluation, and affects. The purpose is to represent the user and adapt the teaching experiences to meet specific learning requirements of the individual. A sample of 43 approaches is stated; where 11 are in progress and 32 are mature.

3.1.1. Student modeling approaches published in 2010

Seventeen approaches are outlined as follows, where 6 are in progress and 11 are mature whose EDM approach profile is presented in Table 12. The incipient approaches series starts as follows: Bian (2010) focus on finding groups of activities in which all students demonstrate similar performance; Bousbia, Labat, Balla, and Rebai (2010) identify students' learning styles from learning indicators; Khodeir, Wanas, Darwish, and Hegazy (2010) build a differential student model based on a probabilistic domain; Rai and Beck (2010) investigate in what class of students benefitted from the computer tutor Mily's World, as well as which students preferred this style of instruction to traditional materials; Rupp, Sweet, and Choi (2010) study learning trajectories by epistemic network analysis; Soundranayagam and Yacef (2010) analyze the order of resource usage and its links with students' learning.

The series of mature approaches begins with Macfadyen and Dawson (2010), id 1. They assert that pedagogically meaningful information can be extracted from LMS-generated student tracking data. Thus, authors investigate which student online activities accurately predict academic achievement. In consequence, they propose regression modeling to incorporate key variables (e.g., total number of discussion messages posted, total number of mail messages sent...). Ping-Feng et al. (2010), id 2, design an improved model of rough set theory. The model analyzes academic achievements of students in Taiwan. Empirical results show the model selects key information from data without predetermining factors and provides accurate rates for inference rules. Guruler, Istanbullu, and Karahasan (2010), id 3, explore the factors having impact on

Table 11
Summary of EDM works, disaggregated in approaches and tools, presented in Section 3 that were published during the period of 2010 to 2013 first quarter.

EDM approaches and tools	Period: 2010 to 2013 1Q			Year 2010			Year 2011			Year 2012			2013 1Q
	Total	Progress	Mature	Total	Progress	Mature	Total	Progress	Mature	Total	Progress	Mature	Total
EDM works	240	83	157	74	37	37	97	30	67	66	16	50	3
EDM approaches	222	72	150	70	34	36	85	23	62	64	15	49	3
1. Student modeling	43	11	32	17	6	11	15	5	10	10	0	10	1
2. Student behavior modeling	48	20	28	18	12	6	15	4	11	13	4	9	2
3. Student performance modeling	46	16	30	0	0	0	22	9	13	24	7	17	0
4. Assessment	45	11	34	18	8	10	18	3	15	9	0	9	0
5. Student support and feedback	21	8	13	12	4	8	3	0	3	6	4	2	0
6. Curriculum, domain knowledge, sequencing, and teachers support	19	6	13	5	4	1	12	2	10	2	0	2	0
EDM tools	18	11	7	4	3	1	12	7	5	2	1	1	0
1. Extraction, learning support, and feature engineering	4	1	3	1	0	1	2	1	1	1	0	1	0
2. Visualization	6	5	1	3	3	0	3	2	1	0	0	0	0
3. Analysis support	8	5	3	0	0	0	7	4	3	1	1	0	0

Table 12
Main traits of the EDM approach profile that depicts student modelling approaches published in 2010.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
1	Macfadyen and Dawson (2010)	Probability	Predictive	Regression	M: social network, analysis (SNA) T: logistic regression, multiple regression	*E*: logistic regression, multiple regression
2	Ping-Feng, Yi-Jia, and Yu-Min (2010)	Machine learning	Predictive	Classification	*M*: rules induction decision tree linear discriminant analysis (LDA) *T*: rough set theory	*A*: PART, C4.5, CART, ID3 *F*: improved rough set model
3	Guruler et al. (2010)	Machine learning	Predictive	Classification	*M*: decision tree	*A*: Microsoft decision tree
4	Arroyo, Mehranian, and Woolf (2010)	Statistic	Predictive	Classification, regression	*T*: Pearson correlation, mean squared error	*E*: statistical
5	Desmarais and Pelczer (2010)	Probability	Predictive	Classification	*M*: item response theory (IRT) *T*: Q-matrix, marginal, probability, covariance	*E*: linear algebra, logistic IRT
6	D'Mello and Graesser (2010)	Probability	Predictive	Regression	*T*: logistic regression	*E*: binary logistic regression
7	Fincham et al. (2010)	Probability, dynamic programming (DP)	Predictive	Classification	*M*: HMM	*A*: Viterbi and naïve Bayes
8	Gong et al. (2010)	Probability Machine learning,	Predictive, descriptive	Clustering, classification	*M*: IBL, Bayes theorem	*A*: k-means, EM
9	Nugent et al. (2010)	Machine learning	Descriptive	Clustering	*M*: IBL	*A*: hierarchical agglomerative clustering, k-means,
10	Pardos and Heffernan (2010a)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: Bayesian knowledge tracing (BKT), EM
11	Yudelson et al. (2010)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: BKT, trust region reflective *E*: mean squared error, linear square fitting

the success of university students. They use the DM tool MUSKUP for classification purposes. The findings unveil: the types of registration to the university and the income levels of the students' family are associated with student success. Arroyo, Mehranian, and Woolf (2010), id 4, developed an effort-based tutoring to model student actions with an integrated view of learner behaviors to represent the real way that students use an ITS.

In another vein, Desmarais and Pelczer (2010), id 5, investigate different methods for generating simulated data. Thus, they compare the predictive performance of a Bayesian student model over real against simulated data for which the parameters are set to reflect those of the real data as closely as possible. D'Mello and Graesser (2010), id6, investigated learners' postural patterns associated with naturally affective states that occur during a tutoring session. They extracted 16 posture-related features that focused on the pressure exerted along with the magnitude and direction of changes in pressure during emotional experiences. Fincham, Anderson, Betts, and Ferris (2010), id 7, infer the students' mental states while they are using an ITS by means of a cognitive model. In addition, functional magnetic resonance imaging is used to predict whether or not students were engaged in problem solving. Gong, Beck, and Heffernan (2010), id 8, model groups of students separately based on their distributional similarities and Dirichlet priors. As a result, a model of parameters provides a plausible picture of student domain knowledge.

Furthermore, Nugent, Dean, and Ayers (2010), id 9, use an empty k-means algorithm to allow for empty clusters and a method based on the Q-matrix to determine efficient starting centers. Combining both items improves the clustering results and allows for an analysis of students' skill set profiles. Pardos and Heffernan (2010a), id 10, pursue to determine when a student has acquired the domain knowledge that a cognitive tutor teaches. Yudelson, Brusilovsky, Mitrovic, and Mathews (2010), id 11, propose a method to improve the user model mapping by using a numerical optimization procedure.

3.1.2. Student modeling approaches published in 2011

Fifteen approaches are highlighted next, where 5 are incipient and 10 are mature whose EDM approach profile is depicted in Table 13. The set of approaches in progress initiates with: Gowda, Baker, Pardos, and Heffernan (2011), who compare single versus ensemble approaches devoted to produce student knowledge model from the ASSISSTments platform. As result, they assert: ensemble approaches produce predictions of student performance 10% better than the best individual student knowledge model; Srinivas et al. (2011) apply a data driven EDM-based methodology to individualize education that embraces: guidance for each student, target instruction for clusters of students, and a mixture of data from several tests with other subjects (e.g., social, economical); Xu and Mostow (2011a) trace multiple sub-skills by logistic regression in a dynamic Bayesian net; Lemmerich, Ifland, and Puppe (2011) identify factors that bias on the overall success of students by subgroup discovery; Yudelson, Pavlik, and Koedinger (2011) compare the Q-matrix and the T-matrix to understand the domain knowledge transference in cognitive models.

The series of mature approaches starts with Narli, Özgen, and Alkan (2011), id 12, who use rough sets theory to identify the relationship between individuals' multiple intelligence and their learning styles. Data is collected from the use of the Multiple Intelligence Inventory for Educators and the Learning Styles Scale tests to mathematics prospective teachers. Desmarais (2011), id 13, analyzes the factors and assumptions under which non-negative matrix factorization can effectively derive the underlying high level skills behind assessment results. Nwaigwe and Koedinger (2011), id 14, investigate the generality of performance of the simple location heuristic and the simple temporal heuristic at predicting student changes in error rate over time. Gong and Beck (2011), id 15, use the same student modeling framework for different evaluations to construct guidance about what student model components are relevant for designing an accurate student model. Koedinger et al. (2011), id 16, apply a conjunctive knowledge

Table 13
Main traits of the EDM approach profile that depicts student modelling approaches published in 2011.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
12	Narli, Özgen, and Alkan (2011)	Machine learning	Predictive	Classification	*M*: decision tree *T*: rough set theory	*A*: generating decision *E*: dependency of attributes
13	Desmarais (2011)	Probability	Descriptive	Clustering	*T*: Q-matrix	*A*: non-negative matrix factorization
14	Nwaigwe and Koedinger (2011)	Probability	Predictive	Regression	*T*: logistic regression	*A*: Bayesian information criterion
15	Gong and Beck (2011)	Probability	Predictive	Regression	*T*: logistic regression	*A*: performance factors analysis
16	Koedinger et al. (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: conjunctive knowledge tracing for fair blame assign. *F*: Bayesian networks
17	Nooraei et al. (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: BKT *F*: Bayesian networks
18	González-Brenes et al. (2011)	DP	Predictive	Classification	*M*: HMM	*A*: hidden conditional random fields
19	Xu and Mostow (2011b)	Probability	Predictive	Regression	*T*: logistic regression	*E*: logistic regression *F*: dynamic Bayesian networks
20	Mostow, Xu, and Munna (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: learning decomposition F: dynamic Bayesian networks
21	Goguadze et al. (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*F*: Bayesian networks

tracing approach to study problem selection thrashing in analysis of log data. Nooraei, Pardos, Heffernan, and Baker (2011), id 17, model students by means of knowledge tracing of a pair of datasets and the most recent 15 data points. As result, they claim: knowledge tracing needs only a small range of data to learn reliable parameters.

As for González-Brenes, Duan, and Mostow (2011), id 18, they apply hidden conditional random fields to predict reading task completion. They formulate tutorial dialogue classification as a sequence classification problem to evaluate dialogue classification. Xu and Mostow (2011b), id 19, use logistic regression over each step's sub-skills in a dynamic Bayesian net to model transition probabilities for the overall knowledge required by the step. Mostow, Xu, and Munna (2011), id 20, design a framework to depict and mechanize the selection of parameters within a model of student learning. So, they implement a heuristic search through a space of alternative parameterizations. Goguadze, Sosnovsky, Isotani, and Mclaren (2011), id 21, evaluate a Bayesian model of student misconceptions, which focuses on presenting and adapting erroneous examples in the decimals domain.

3.1.3. Student modeling approaches published from 2012 up to 2013 1st Qtr

Eleven mature approaches are introduced in this group, where ten are published in 2012 and just one in 2013. The EDM approach profile that characterizes the approaches is outlined in Table 14, whereas the series of approaches is described thereafter.

Holzhüter, Frosch-Wilke, and Klein (2012), id 22, pursue solving a couple of issues: how can learning processes be optimized using process models and rule-based control? How can process models be generated based on the learning style concept? So, they propose a method of learner modeling by means of combining the process mining and the learning style approach as a method of learner modeling. Yanto, Herawan, Herawan, and Deris (2012), id 23, demonstrate the applicability of a variable precision rough set model for clustering students who suffer from anxiety. The approach is based on the mean of accuracy of approximation using variable precision of attributes. Rupp et al. (2012), id 24, build evidence rules and measurement models within the evidence model of the evidence-centered design framework in the context of the Cisco Networking Academy Digital learning environment. Sparks, Patton, and Ganschow (2012), id 25, examine achievement, intelligence, aptitude, and proficiency profiles of students by cluster analysis to determine whether distinct cognitive and achievement of more and less successful learners emerges. Rus, Moldovan, Niraula, and Graesser (2012), id 26, address the discovery of speech act categories in dialogue-based multi-party educational games. They shape a data-driven method to discover speech act taxonomies.

As regard with Trivedi, Pardos, Sárközy, and Heffernan (2012), id 27, they conceptualize a bagging strategy with co-clustering in order to predict results of out-of-tutor performance of students. González-Brenes and Mostow (2012), id 28, propose a unified model, called Dynamic Cognitive Tracing, to explain student learning in terms of skill mastery over time by learning the cognitive model and the student model jointly. Rau and Scheines (2012), id 29, mine log data on error-rate, hint-use, and time-spent obtained from Fractions. They compare the achieved learning from multiple graphical representations of Fractions versus the acquired learning from a single graphical representation. Baker et al. (2012), id 30, tailor models to detect student engaged concentration, confusion, frustration, and boredom solely from students' interactions within Algebra. The detectors operate solely on the information available through students' semantic actions within the interface. Eagle, Johnson, and Barnes (2012), id 31, design a data structure for the analysis of interaction-data collected from open problem solving environments. Such a data is mined through network sciences techniques. Nandeshwar, Menzies, and Nelson (2013), id 32, are interested in improving learning predictors for student retention. They explore many learning methods, carefully select traits, and evaluate the efficacy of the learned theory by its median and the variance during the performance.

3.2. Student behavior modeling

Student modeling devoted to characterize *behavior* is one of the preferred targets of EDM approaches. Diverse traits of behavior are the subject of modeling, such as: gambling, guessing, sleeping, inquiring, requesting help, willingness to collaborate, time series of access and response, and many more targets. The purpose is to describe or predict particular pattern behaviors in order to adapt

Table 14
Main traits of the EDM approach profile that depicts student modelling approaches published from 2012 up to 2013 1st Qtr.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: model
22	Holzhüter et al. (2012)	Machine learning	Descriptive	Association rules	*M*: decision tree	*A*: heuristic miner, ProM
23	Yanto et al. (2012)	Machine learning	Descriptive	Clustering	*T*: rough sets	*A*: variable precision rough set
24	Rupp et al. (2012)	Probability, statistic	Predictive	Classification	*M*: IRT, diagnostic classification models, Bayes theorem,	*F*: Bayesian networks
25	Sparks et al. (2012)	Machine learning	Descriptive	Clustering	*M*: IBL, cluster analysis	*A*: k-means
26	Rus et al. (2012)	Probability, machine learning	Descriptive	Clustering	*M*: IBL, Bayes theorem	*A*: k-means
27	Trivedi et al. (2012)	Probability, machine learning	Descriptive	Clustering	*M*: co-clustering bipartite graph partitioning problem	*A*: bipartite spectral graph partitioning, prediction model
			Predictive	Regression	*T*: linear regression	
28	González-Brenes and Mostow (2012)	Probability	Descriptive	Clustering, classification	*M*: Bayes theorem	*A*: EM, junction tree
			Predictive			*F*: Bayesian networks
29	Rau and Scheines (2012)	Probability	Predictive	Regression	*T*: linear regression	*A*: stepwise linear regression
30	Baker et al. (2012)	Probability, machine learning	Predictive	Classification, regression	*M*: Bayesian theorem, decision tree *T*: linear regression	*A*: stepwise linear regression Naïve Bayes, J48
31	Eagle et al. (2012)	Artificial intelligence	Descriptive	Clustering	*T*: interactions network	*A*: Girvan-Newman, Bellman backup
32	Nandeshwar et al. (2013)	Probability, machine learning	Predictive	Classification	*M*: Bayes theorem, decision tree, neural networks, rules induction	*A*: One-R, C4.5, NaiveBayes, radial basis function *F*: Bayesian networks

the system to the users' tendencies. This review presents 48 approaches, where 20 are incipient and 28 are mature, as follows:

3.2.1. Student behavior modeling approaches published in 2010

Eighteen approaches are introduced next, where 12 are in progress and 6 are mature. The EDM approach profile of the mature approaches is given in Table 15 and their profile is stated thereafter. As the first part of the incipient works, a briefing of the Cup 2010 Workshop Knowledge Discovery in Educational Data is reported. During the event, the teams designed a model from students' behavior and then predicted future performance. The source given to the competitors was log data from Algebra 2008–2009 and Bridge to Algebra 2008–2009. A briefing of nine approaches is presented as follows:

Yu et al. (2010) expand features by binarization and discretization techniques. The resulting sparse feature sets are trained by L1-regularized logistic regression. Next, the features are condensed by statistical techniques and random forest. Finally, the results are combined by regularized linear regression; Toscher and Jahrer (2010) predict student's ability to answer questions correctly, based on historical results. They use an ensemble of collaborative filtering techniques; Pardos and Heffernan (2010b) extract features and predict students' outcomes by means of Bayesian HMM and bagged decision tree methods; Shen et al. (2010) tailor a framework to handle asymmetric training and test sets as well as non-atomic and variable-length attributes; Gandhi and Aggarwal (2010) apply the Rasch model technique to capture the effects of student level proficiency and steps' level difficulty. Moreover, they use a hybrid ensemble of logistic regression models to produce predictive results.

Furthermore, Wijaya and Prasetyo (2010) compute the probability about student's understanding of a particular problem by an exponential moving average to give more weighting to the recent results; Tabandeh and Sami (2010) assert: despite using only three features of 22 features to model learners, they reach acceptable results by regular decision trees and regression algorithms; Perez-Mendoza, Rubens, and Okamoto (2010) sketch a hierarchical aggregation prediction method to achieve hierarchical aggregation of data and feature selection; Liu and Xing (2010) aim at developing a predictive model of student's behavior by an ensemble

Table 15
Main traits of the EDM work approach that depicts student behaviour modelling approaches published in 2010.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *M*: model
33	Baker and Gowda (2010)	Statistic	Descriptive	Clustering	*T*: frequencies, variability	*E*: statistical
34	Montalvo, Baker, Sao-Pedro, Nakama, and Gobert (2010)	Machine learning	Predictive	Classification	*M*: decision tree	*A*: J48
35	Sao-Pedro et al. (2010)	Machine learning	Predictive	Classification	*M*: decision tree	*A*: J48
36	Romero, Romero, Luna, and Ventura (2010)	Machine learning	Descriptive	Association rules	*T*: rare association rule mining	*A*: four Apriori: frequent, infrequent, inverse, rare
37	Shanabrook et al. (2010)	Probability	Descriptive	Sequential pattern	*T*: projection-based search time-based motif	*A*: random projection multivariate motif discovery
38	Shih et al. (2010)	DP	Descriptive	Clustering	*M*: HMM	*A*: stepwise-HMM-cluster

approach composed of creation of sampled sets, generation of base models, and selection of base models to be aggregated for obtaining the final ensemble model.

In addition to the alluded event, three incipient approaches are stated next: Forsyth et al. (2010) investigate the conditions in which the length of the students' contributions is correlated with learning; González-Brenes and Mostow (2010) define a data-driven model to predict task completion; Hershkovitz and Nachmias (2010) identify individuals' over-time patterns of online activity in LMS.

As for the mature approaches, the first is made by Baker and Gowda (2010), id 33, who investigate the behavior of students, who used the Geometry ITS in urban, rural, and suburban schools. They find that learners in the urban school go off-task and are careless significantly more than students in the rural and suburban schools. Montalvo et al. (2010), id 34, detect student metacognitive planning processes. So, they develop detectors for students' planning of experiments by tracing time spent looking at data tables. Sao-Pedro, Baker, Montalvo, Nakama, and Gobert (2010), id 35, also detect two forms of students' systematic data collection behavior, control of variables strategy, and hypothesis testing, which are shown within a virtual phase change in Science ASSISTments.

What is more, Romero, Romero, Luna, and Ventura (2010), id 36, explore rare/infrequent learners' behaviors when using a LMS. So, they implement several Apriori algorithms to discover rare association rules from log data. They evaluate the relation/influence between the online activities and the final marks obtained by the students. Shanabrook, Cooper, Woolf, and Arroyo (2010), id 37, examine student interaction with the Wayang Outpost ITS during problem solving. They discover student behavior patterns by mining student actions tracked as logs during tutor sessions. Shih, Koedinger, and Scheines (2010), id 38, propose the stepwise-HMM-cluster algorithm for HMM. It discovers student learning tactics while incorporating student-level outcome data, constraining the results to interpretable models.

3.2.2. Student behavior modeling approaches published in 2011

Fifteen approaches are introduced in this subsection; where four are incipient and eleven mature whose EDM approach profile is presented in Table 16. The first incipient approach corresponds to: Fancsali (2011) searches for variable constructions from raw student messaging data in an online forum; Merceron (2011) studies whether a core group of students emerges that keep using the resources or whether, on the contrary, students are eclectic in their choice, consulting resources randomly; Zorrilla, García-Saiz, and Balcázar (2011) compare several algorithms for association rules on educational datasets to test whether a Yacaree-based approach is useful for the EDM; Inventado, Legaspi, Suarez, and Numao (2011) hypothesize observing whether affect will help to understand the transitions between learning and non-learning activities when students learn online.

The collection of mature approaches begins with Levy and Wilensky (2011), id 39, who investigate students' inquiry actions in three models of complex chemical systems while their goal is to construct an equation relating physical variables of the system. They explore whether and how students adapt to different behaviors of the system. Huei-Tse (2011), id 40, explores the learning process of adopting collaborative online instructional discussion activities for the purpose of problem-solving using situated scenarios. Köck and Paramythis (2011), id 41, represent learners' problem solving activity sequences to detect predefined and problem solving styles. They analyze learner behavior along known learning dimensions to semi-automatically discover learning dimensions and concrete problem solving patterns. Muldner, Burleson, Van de Sande, and VanLehn (2011), id 42, make the question: What is a better predictor of gaming, problem or student? So, they develop a gaming detector for automatically labeling the log data, and apply Bayesian network. As result, they find the student is a better predictor of gaming than problem.

In addition, Anaya and Boticario (2011a), id 43, build a domain-independent modeling method of collaborative learning based on DM that helps to clarify which user-modeling issues need to be

Table 16
Main traits of the EDM work approach that depicts student behaviour modelling approaches published in 2011.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
39	Levy and Wilensky (2011)	Probability	Predictive	Regression	*T*: logistic regression	*E*: logistic regression
40	Huei-Tse (2011)	Machine learning	Descriptive	Clustering	*M*: IBL *T*: hierarchical clustering	*A*: k-means
41	Köck and Paramythis (2011)	DP	Descriptive	Clustering	*M*: discrete Markov model *T*: sequential pattern	*A*: linear discriminant analysis
42	Muldner et al. (2011)	Probability	Predictive	Regression	*M*: Bayes theorem *T*: linear regression	*E*: linear regression, frequencies *F*: dynamic Bayesian network
43	Anaya and Boticario (2011a)	Probability Machine learning	Descriptive	Clustering	*M*: decision tree, Bayes theorem	*A*: EM, J48, REPTree, simple Cart, NaïveBayes
44	Hershkovitz and Nachmias (2011)	Statistic, machine learning	Descriptive	Clustering	*M*: decision tree	*A*: CHAID *E*: descriptive statistical
45	Martinez, Yacef, Kay, Al-Qaraghuli, and Kharrufa (2011)	Probability	Descriptive	Sequential pattern	*T*: hierarchical clustering	*A*: hierarchical agglomerative clustering
46	Qiu et al. (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: BKT-forget, BKT-slip *F*: Bayesian network
47	Kardan and Conati (2011)	Machine learning	Descriptive, predictive	Clustering, association rules, classification	*M*: IBL	*A*: k-means, genetic k-means
48	Cobo et al. (2011)	Probability	Descriptive	Clustering	*T*: hierarchical clustering	*A*: hierarchical agglomerative clustering
49	Ivancevic et al. (2011)	Machine learning	Descriptive	Association rules	*M*: IBL, support vector machines	*A*: k-means, anomaly detection

considered. Hershkovitz and Nachmias (2011), id 44, mine log files of 58 Moodle websites to identify the degree of persistence of learners' online activity. As result, they assert: 42% of learners persist or accelerate their activity towards the end of the semester, against 46% who decelerated or quit their activity. Martinez, Yacef, Kay, Al-Qaraghuli, and Kharrufa (2011), id 45, exploit the log traces of the Digital Misteries CBES to extract patterns of activity for unveiling the strategies followed by groups of learners. One year after, Martinez, Yacef, and Kay (2012), id 58, present a method to capture, exploit, and mine the digital footprints of students working face-to-face to build a concept map at an interactive tabletop.

Qiu, Qi, Lu, Pardos, and Heffernan (2011), id 46, work on knowledge tracing predictions on student responses where more than one day had elapsed since the previous response and find knowledge tracing consistently over the predicted data points. Kardan and Conati (2011), id 47, tailor a user modeling framework that relies on interaction logs to identify types of learners, as well as their characteristic interaction behavior and how the behaviors relate to learning. Cobo et al. (2011), id 48, identify what behaviors patterns are adopted by students in online forums. Ivancevic, Celikovic, and Lukovic (2011), id 49, examine relationships between student assessment results and student choices of seating locations in a lab.

3.2.3. Student behavior modeling approaches published from 2012 up to 2013 1st Qtr

Fifteen approaches are summarized in this subsection, where thirteen are published in 2012 and two in 2013. Four represent incipient approaches and eleven are mature, but one of them is presented beforehand. The EDM approach profile of the mature works is stated in Table 17. The series of approaches in progress commences with Keshtkar, Morgan, and Graesser. (2012), who investigate the dynamics and linguistic features of multi-party chat in the context of an online educational game; Rafferty, Lamar, and Griffiths (2012) outline a framework for automatically inferring a student's underlying beliefs from a set of observed actions, which relies on modeling how student actions follow from beliefs about the effects of those actions; Merceron et al. (2012) study learning paths in a non-personalizing e-Learning environment; Fancsali (2012) sketches a method to simultaneously search for student-level variables constructed from log data and graphical causal models.

The first instance of mature approaches corresponds to Sweet and Rupp (2012), id 50, who demonstrate how the evidence-centered design framework provides critical support for tailoring simulation studies to investigate statistical methods within a defined methodological domain like games-based assessment. Antonenko, Toy, and Niederhauser (2012), id 51, mine click-stream server-log data that reflects student use of online learning environments. They apply cluster analysis to analyze characteristics of learning behavior while learners engage in a problem-solving activity. Patarapichayatham, Kamata, and Kanjanawasee (2012), id 52, evaluate the impact of model selection strategies. As result, they find the Bayesian information criterion strategy tends to choose incomplete models more often than other strategies and leads to more biased parameter estimations. Bouchet, Azevedo, Kinnebrew, and Biswas (2012), id 53, examine trace data to identify distinguishing patterns of behavior in an analysis of students learning about a science topic by means of an ITS based on agent that fosters self-regulated learning.

As for Peckham and McCalla (2012), id 54, determine positive and negative cognitive skill sets with respect to reading comprehension by multidimensional k-means clustering combined with Bloom's Taxonomy. Bayer, Bydzovská, Géryk, Obsıvac, and Popelínský (2012), id 55, predict drop-outs and school failures when student data has been enriched with data derived from students' social behavior. The data unveils social dependencies gathered from e-mail and discussion board conversations. Sabourin, Mott, and Lester (2012), id 56, predict student self-regulation learning

Table 17
Main traits of the EDM approach profile that depicts student behaviour modelling approaches published from 2012 up to 2013 1st Qtr.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
50	Sweet and Rupp (2012)	Statistic, probability	Descriptive	Clustering	*M*: SNA, IRT, epistemic network analysis	*E*: statistical
51	Antonenko et al. (2012)	Statistic	Descriptive	Clustering	*M*: IBL *T*: correlation analysis	*A*: cluster analysis, k-means
52	Patarapichayatham et al. (2012)	Statistic, probability	Descriptive	Clustering	*M*: IRT *T*: factor analysis exploratory, Rasch model	*A*: factor analysis performance *E*: maximum likelihood estimate *F*: Akaike information criterion, Bayesian information criterion, differential item functioning, likelihood ratio tests
53	Bouchet et al. (2012)	Machine learning	Descriptive	Clustering	*M*: Bayes theorem, differential sequencing	*A*: EM
54	Peckham and McCalla (2012)	Machine learning	Descriptive	Clustering	*M*: IBL	*A*: k-means
55	Bayer et al. (2012)	Probability, machine learning	Predictive	Classification	*M*: Bayes theorem, decision tree induction rules, entropy,	*A*: J48, PART, NaiveBayes InfoGainAttributeEva,
56	Sabourin et al. (2012)	Probability	Predictive	Regression	*T*: logistic regression	*A*: stepwise logistic regression
57	McCuaig and Baldwin (2012)	Machine learning	Predictive	Classification	*M*: decision tree	*A*: decision tree recursive partitioning
58	Martinez, Yacef, and Kay (2012)	Probability	Descriptive	Sequential pattern	*T*: hierarchical clustering	*A*: hierarchical agglomerative clustering
59	He (2013)	Statistic	Descriptive	Clustering	*T*: correlation analysis	*A*: cluster analysis *E*: descriptive statistical
60	Malmberg et al. (2013)	Machine learning	Descriptive	Clustering	*M*: IBL	*A*: k-means

capabilities. So, they classify students into self-regulation learning-use categories based on evidence of goal-setting and monitoring activities. McCuaig and Baldwin (2012), id 57, assert that source log data produced by conventional LMS could be mined to predict the students' success or failure without requiring the results of formal assessments.

Concerning the pair of mature works published in 2013, the first corresponds to He (2013), id 59. He analyses online questions and chat messages recorded by a live video streaming. The study identifies discrepancies and similarities in the students' patterns and themes of participation between student–instructor interaction, as well as student–students interaction or peer interaction. Concerning Malmberg, Järvenoja, and Järvela (2013), id 60, they investigate the learning patterns that emerge in learning situations that are favorable and challenging. Thereby, they identify differences between high and low achieving students' strategic actions in varying learning situations. The outcomes unveil that both kind of students adopted similar strategies in favorable learning situations.

3.3. Student performance modeling

Student modeling oriented to represent and anticipate *performance* is one of the favorite targets of EDM approaches. Many indicators of performance are worthy to be modeled, such as: efficiency, evaluation, achievement, competence, resource consuming, elapsed time, correctness, deficiencies, etc. The goal is to estimate how well the learner is or will be able to accomplish a given task, reach a specific learning goal, or appropriately respond to a particular learning situation. The survey embraces 46 approaches composed of 16 incipient and 30 mature approaches, whose profile is stated as follows.

3.3.1. Student performance modeling approaches published in 2011

Twenty two approaches are published in 2011, where 9 are in progress and 13 are mature. Table 18 shows the EDM approach profile of the mature approaches. The first incipient work corresponds to Thai-Nghe, Drumond, Horvath, and Schmidt-Thieme (2011). They predict student performance by exploiting multiple relationships between student-tasks-skills by multi-relational matrix factorization methods; Wang, Kehrer, Pardos, and Heffernan (2011) shape the Tabling method of predicting student performance by calculating the expected outcome of students with the same sequence of responses; Xiong, Pardos, and Heffernan (2011) determine the utility of students' response time in performance prediction, so they make experimental observations and analysis on the response time data in the ASSISTments dataset; Hershkovitz, Baker, Gobert, and Wixon (2011) study the relationship between goal orientation within ASSISTments, and the manifestation of carelessness over consecutive trials.

In addition, Kabakchieva, Stefanova, and Kisimov (2011) seek patterns to predict student performance at the university based on their personal and pre-university traits; Wang and Heffernan

Table 18
Main traits of the EDM approach profile that depicts student performance modelling approaches published in 2011.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
61	Wang and Liao (2011)	Neural networks	Predictive	Classification	*M*: back propagation	*A*: gradient-descent
62	Thai-Nghe et al. (2011b)	Probability	Predictive	Classification	*M*: factorization model *T*: factorization, forecasting	*A*: tensor factorization forecasting
63	Li et al. (2011)	Machine learning	Descriptive	Associate rules	*T*: production rules	*A*: deep feature learner
64	Chi, Koedinger, Gordon, Jordan, VanLehn (2011)	Probability	Predictive	Classification	*M*: instructional factors analysis model	*E*: instructional factors analysis model
65	Mostow, González-Brenes, and Tan (2011)	Artificial intelligence	Descriptive	Clustering	*M*: IBL	*A*: k-means
66	Baker et al. (2011)	Probability, artificial intelligence	Predictive	Regression	*M*: feature engineering *T*: linear regression	*E*: linear regression
67	Pardos et al. (2011)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: BKT-contextual guess, BKT-contextual slip, BKT-prior per student, performance-factor analysis *F*: Bayesian networks
68	Gowda, Rowe, Baker, Chi, and Koedinger (2011)	Probability, artificial intelligence	Predictive	Regression	*M*: Bayes theorem, feature engineering *T*: linear regression	*A*: BKT-brute force
69	Akcapinar et al. (2011)	Machine learning	Predictive	Regression	*M*: decision tree	*A*: random forest regression
70	Marquez-Vera et al. (2011)	Machine learning	Predictive	Classification	*M*: decision tree rules induction	*A*: JRip, NNge, OneR, Prism, Ridor, simpleCart, ADTree, randomTree, REPTree, J48
71	Pavlik and Wu (2011)	Probability	Predictive	Classification	*M*: dynamical systems	*E*: additive factor model
72	Rai and Beck (2011)	Statistic	Descriptive	Correlation analysis	*M*: causal model *T*: covariance	*A*: PC search
73	Zafra et al. (2011)	Machine learning, probability	Predictive	Classification	*M*: Bayes theorem, decision tree, neural networks, rules induction, support vector machine	*A*: NaiveBayes, J48, ZeroR, multilayer perceptron, sequential minimal optimization

(2011) predict students' performance, particularly the issues concerning with forgetting and relearning; Zimmermann, Brodersen, Pellet, August, and Buhmann (2011) analyze the statistical relationship between B.Sc. and M.Sc. achievements using a dataset that is not subject to an admission-induced selection bias; Sudol-Delyser and Steinhart (2011) depict the students' learning progression through the activities, and compare student performance on common tutoring questions; Jarusek and Pelánek (2011) offer a problem response theory to predict how much time a student needs to solve a problem.

The series of mature approaches initiates with Wang and Liao (2011), id 61, who explore the recent learning performance of students to predict future performance in learning English. Thai-Nghe, Horvath, and Schmidt-Thieme (2011), id 62, apply the sequential effect (i.e., domain knowledge improves and accumulates over time) for forecasting student performance based on tensor factorization forecasting. Li et al. (2011), id 63, design SimStudent, a machine learning agent to automatically discover skill knowledge acquired to model the student. Chi, Koedinger, Gordon, Jordan, and VanLehn (2011), id 64, propose the Instructional Factors Analysis Model to depict student's performance when multiple types of instructional interventions are involved and some may not generate a direct observation of students' performance.

As well as, Mostow, González-Brenes, and Tan (2011), id 65, develop the Automatic Classifier of Relational Data system to perform the feature engineering by training classifiers directly on a relational database of events logged by a tutor. Their system also learns a classifier to predict whether a child finishes reading a story in LISTEN's Reading Tutor. Baker, Gowda, and Corbett (2011), id 66, build a detector to predict a student's later performance on a paper test of preparation for future learning. This post-test demands learning material to solve problems involving skills that are related, but different to the skills studied in the ITS. Pardos, Gowda, Baker, and Heffernan (2011), id 67, achieve ensembling at the post-test level to check if the approach produces better prediction of post-test scores within the context of an ITS. Gowda, Rowe, Baker, Chi, and Koedinger (2011), id 68, evaluate a set of engineered features that quantify skill difficulty and related skill-level constructs in terms of their ability to improve models of guessing, slipping, and detecting moment-by-moment learning.

What is more, Akcapinar, Cosgun, and Altun (2011), id 69, predict users' perceived disorientation by using random forest regression. Marquez-Vera, Romero, and Ventura (2011), id 70, estimate final student performance and anticipate which students might fail. Thus, they develop two approaches to resolve the problem of classifying unbalanced data by rebalancing data and using cost sensitive classification. Pavlik Jr. and Wu (2011), id 71, develop a dynamical systems model of self-regulated learning to explain dynamic relationships between student-engagement constructs and performance during learning. Rai and Beck (2011), id 72, apply a

Table 19
Main traits of the EDM approach profile that depicts student performance modelling approaches published in 2012.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
74	Schoor and Bannert (2012)	Statistic, soft computing	Descriptive	Clustering	*M*: fuzzy miner *T*: correlation analysis	*A*: descriptive statistic
75	Kerr and d'Chung (2012)	Statistic	Descriptive	Clustering	*T*: correlation analysis	*A*: clustering analysis, fuzzy cluster, hard cluster
76	Koedinger et al. (2012)	Statistic	Predictive	Classification	*M*: additive factors model	*A*: learning factors analysis
77	Xu and Mostow (2012)	Probability	Predictive	Regression	*T*: logistic regression	*A*: EM *E*: logistic regression, conjunctive knowledge tracing *F*: dynamic Bayes networks
78	Goldin et al. (2012)	Probability	Predictive	Regression	*T*: logistic regression	*E*: ProfHelp, ProfHelp-ID
79	Beheshti et al. (2012)	Probability	Predictive	Classification	*T*: factorization, singular value decomposition	*A*: wrapper selection feature
80	Bergner et al. (2012)	Machine learning	Predictive	Classification	*M*: IRT, *T*: collaborative filtering	*E*: binary classifier based on logistic function
81	Lee and Brunskill (2012)	Probability	Predictive	Classification	*M*: HMM, Bayes theorem	*A*: BKT, EM, ExpOppNeed
82	Rau and Pardos (2012)	Probability	Predictive	Classification	*M*: HMM, Bayes theorem	*A*: BKT, EM, Bayes networks
83	Forsyth et al. (2012)	Probability	Predictive	Regression	*T*: logistic regression	*A*: stepwise logistic regression
84	Wang and Heffernan (2012)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: EM *F*: Bayes networks
85	Molina et al. (2012)	Machine learning, probability	Predictive	Classification	*M*: Bayes theorem, decision tree, neural networks, rules induction	*A*: NaiveBayes, J48, PART, multilayer perceptron
86	Yoo and Cho (2012)	Machine learning	Descriptive	Asociación rules	*M*: rules induction	*A*: frequent subgraph discovery apriori
87	Yudelson and Brunskill (2012)	Probability	Predictive	Regression	*T*: logistic regression factor analysis	*E*: logistic regression based on contextual factor analysis
88	Pardos et al. (2012)	Probability	Predictive	Classification, regression	*M*: decision tree *T*: linear regression	*A*: random forest, stepwise linear regression
89	Stamper et al. (2012)	Probability	Predictive	Classification	*M*: Bayes theorem	*A*: BKT adaptive sequencing
90	Wang and Beck (2012)	Probability	Predictive	Regression	*T*: logistic regression	*A*: stepwise logistic regression

causal modeling approach to analyze and explore the data from a game-like math tutor, called Monkey's Revenge. Zafra, Romero, and Ventura (2011), id 73. It is based on multiple instances learning to predict student's performance and to improve the obtained results using single instance learning.

3.3.2. Student performance modeling approaches published in 2012

Twenty five approaches are outlined in this subsection, where 7 are incipient and 17 are mature whose EDM approach profile is presented in Table 19. The collection of approaches in progress begins with García-Saiz and Zorrilla (2012), who sketch a method to eliminate outliers as a previous step to build a classifier; Campagni, Merlini, and Sprugnoli (2012) analyze the path of how a student implemented exams over the degree e-Learning time with the goal to understand how this order affects the performance of the students in terms of graduation time and final grade; Warnakulasooriya and Galen (2012) demonstrate how students' response patterns can be quantified both globally and locally using the fractal dimension concept; Chaturvedi and Ezeife (2012) propose an approach to integrate mined concept examples at different difficulty levels with the Bayesian networks in order to influence student learning; Sun (2012) builds an approach to find the most dependent test items in students' response data by adopting the entropy concept from information theory; Tan (2012) studies a class of fit-to-model statistics for quantifying the evidence used in learning Bayesian student ability estimation; Crespo and Antunes (2012) advise quantifying the students performance in teamwork by making use of the most effective techniques for social networks analysis.

The first mature approach of the sample corresponds Schoor and Bannert (2012), id 74, who study sequences of social regulatory processes (i.e., individual and collaborative activities of analyzing, planning... aspects) during collaborative sessions and their relationship to group performance. Kerr and d'Chung (2012), id 75, are interested in identifying key features of student performance as a relevant step of the assessment cycle of evidence-centered design. Koedinger, McLaughlin, and Stamper (2012), id 76, work on automated improvement of student models that leverages a crowd educational repository. Xu and Mostow (2012), id 77, model knowledge tracing, fit its parameters, predict performance, and update sub-skill estimates as an attempt to solve how update estimates of multiple sub-skills underlie a single observable step. Goldin, Koedinger, and Aleven (2012), id 78, propose logistic regression models, ProfHelp and ProfHelp-ID, to represent the effect of hints on performance at the same step when the help is provided.

As for Beheshti, Desmarais, and Naceur (2012), id 79, they find the skills behind a set of exercise and question items by determining the number of dominant latent skills that are influential enough to define the item success. Bergner et al. (2012), id 80, apply IRT and model-based collaborative filtering to find parameters for students and items that are combined to predict student performance on an item by item basis. Lee and Brunskill (2012), id 81, use the Knowledge Tracing framework and discover that: when fitting parameters to individual students, there is a variation among the individual's parameters. Rau and Pardos (2012), id 82, apply knowledge tracing (KT) to augment the results obtained from an experiment devoted to investigate the effects of practice schedules using an ITS for fractions. Forsyth et al. (2012), id 83, discover different trajectories of learning within eleven core concepts by evaluating three main constructs (e.g., discrimination, generation, and time on task) represented by key logged measures. Wang and Heffernan (2012), id 84, determine whether the information of student first response time of a question can be leveraged into a KT model and improve the KT prediction accuracy.

As for Molina, Luna, Romero, and Ventura (2012), id 85, they develop a work concerned with meta-learning for tuning parameters. Yoo and Cho (2012), id 86, study the feasibility of the analysis of concept maps and investigate the possibility of using concept maps as a research tool to understand college student's learning. Yudelson and Brunskill (2012), id 87, prescribe activities that maximize the knowledge acquisition as evaluated by expected post-test success. Pardos, Wang, and Trivedi (2012), id 88, provide an analysis of the error differences impact on student test score prediction. Stamper et al. (2012), id 89, design the Super Experiment Framework that guides how internet-scale experiments can inform and be informed by classroom and lab experiments. Wang and Beck (2012), id 90, predict student performance after a delay to determine if, and when, the student will retain the acquired knowledge.

3.4. Assessment

The supervision and evaluation of learners' domain knowledge acquisition, skills development, and achieved outcomes, as well as reflection, inquiring, and sentiments are essential subjects to be taken into account by CBES. The purpose is to differentiate student proficiency at the finer grained level through static and dynamic testing, as well as online and offline assessment. This survey offers a set of 45 citations made up of 11 approaches in progress and 34 mature whose profile is given in this subsection.

3.4.1. Assessment approaches published in 2010

Eighteen approaches are introduced next, where 8 are in progress and 10 are mature. The EDM approach profile of the mature is revealed in Table 20; whereas the series of incipient approaches starts with Buldua and Üçgüna (2010), who apply association rules to find recurrent failures such as: where students who are unsuccessful in numeral courses become unsuccessful again in those courses one year later; Aleahmad, Aleven, and Kraut (2010) rate crowd-sourced examples to determine which are worthy of presenting to students; Bachmann, Gobert, and Beck (2010) assess students' inquiry processes within Microworlds; Goldstein, Baker, and Heffernan (2010) design a model to infer the probability that a student learned a skill at a step during the problem-solving process.

As for Pavlik (2010), he tailors data reduction methods to understand learning hypotheses by the manipulation of specific variables to improve responses on a posttest; Romero, Ventura, Vasilyeva, and Pechenizkiy (2010) apply class association rule from students' test data to discover relationships; Songmuang and Ueno (2010) build test forms to satisfy the common test constraints and ensure that all forms have equivalent qualities; Wang, Heffernan, and Beck (2010) model student performance with continuous partial credit, which is assigned according to the details mined from the student responses.

Concerning the series of mature approaches, the first is made by Sohn and Ju (2010), id 91, who use conjoint analysis to assign weights to four sources of candidates data (e.g., test, record, essay, and interview) to effectively recruit aspirants who have a high quality. Kuncel, Wee, Serafin, and Hezlett (2010), id 92, investigate whether the Graduate Record Examination predicts the performance of students in Master's programs as well as the performance of doctoral students. They find that such a test predicts first year grade point average, graduate, and faculty ratings as well as for both master and doctoral students. France, Finney, and Swerdzewski (2010), id 93, study learners' group and member attachment to their university by the University Attachment Scale. They find a two-factor model is better than a one-factor model providing evidence of a distinction between university attachment and member attachment.

Table 20
Main traits of the EDM approach profile that depicts assessment approaches published in 2010.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
91	Sohn & Ju, 2010	Statistic	Descriptive	Clustering	*M*: conjoint analysis *T*: descriptive statistic	*E*: quality function deployment, statistical
92	Kuncel et al. (2010)	Statistic	Predictive	Classification	*T*: meta-analysis	*E*: descriptive statistic
93	France et al. (2010)	Statistic	Descriptive	Correlation analysis	*M*: factor model, two factor model	*E*: descriptive statistic, factor model
94	Cetintas et al. (2010)	Probability, machine learning	Predictive	Classification	*M*: joint probabilistic, classification model, support vector machine	*A*: linear support vector machines, improved classification model support vector machines
95	Feng and Heffernan (2010)	Probability	Predictive	Regression	*T*: linear regression	*A*: stepwise linear regression
96	Howard et al. (2010)	Statistic, probability, DP	Descriptive	Clustering	*M*: Petri nets, HMM	*A*: frequencies techniques; Petri nets-based. HMM–based *E*: statistical
97	Jeong et al. (2010)	DP	Descriptive	Clustering	*M*: HMM	*A*: HMM–based
98	Xiong et al. (2010)	Machine learning	Predictive	Classification	*M*: decision tree	*A*: J48, k-means *E*: statistical
99	Falakmasir and Habibi (2010)	Machine learning	Predictive	Classification	*M*: decision tree *T*: gain ratio attribute evaluation	*A*: C4.5, attribute evaluation
100	Rajibussalim (2010)	Statistic, machine learning	Descriptive, predictive	Correlation analysis, clustering, classification	*M*: decision trees, statistical analysis, IBL	*A*: BKT, EM, ExpOppNeed

In another vein, Cetintas et al. (2010), id 94, estimate the difficulty level of math word problems by pondering the relevance of sentences through joint probability of classification decisions for all sentences of a math word problem. Feng and Heffernan (2010), id 95, apply online metrics for dynamic testing that measures student accuracy, speed, attempts, and help-seeking behaviors to predict student state test scores when they use ASSISTment. Howard, Johnson, and Neitzel (2010), id 96, examine how learners use a structured inquiry cycle strategy by process analysis, a suitable technique to achieve a formative assessment data from the visited modules of an e-Learning system.

Regarding the work developed by Jeong, Biswas, Johnson, and Howard (2010), id 97, it applies HMM to analyze students' activity sequences and map them onto potential learning behaviors. So, they examine ways to depict the behaviors of students, and determine whether the high-performing students have learning behaviors that are distinct from the low performers. Xiong, Litman, and Schunn (2010), id 98, propose an evaluation system that generates assessment on reviewers' evaluation skills regarding the issue of problem localization. Falakmasir and Habibi (2010), 99, analyze the web usage records of students' activities in a LMS to rank the learning activities based on their impact on the performance of students in final exams. Rajibussalim. (2010), id 100, evaluates the effectiveness of EDM for the extraction of knowledge about the impact of reflection on learning, gaining knowledge about students' learning behavior, and identifying which behavioral patterns lead to positive or negative outcomes.

3.4.2. Assessment approaches published in 2011

Eighteen assessment approaches are published in 2011, where 3 are incipient and 15 mature whose EDM approach profile is given in Table 21. Regarding the approaches in progress, the first is fulfilled by Anaya and Boticario (2011b).They infer collaborative significant student's characteristics in terms of activity and initiative, as well as student acknowledgment of fellow-students; He, Veldkamp, and Westerhof (2011) build an automatic computerized coding framework to identify the characteristics of redemption and contamination in life narratives written by students; Von-Davier (2011) applies quality control and DM tools from text analysis in order to scale scores and other variables of an assessment.

The first mature approach corresponds to Rad, Naderi, and Soltani (2011), id 101, who tackle the problem of clustering and ranking university majors in Iran. Based on eight criteria, they cluster 177 university majors according their similarities and differences. Hsu, Chou, and Chang (2011), id 102, conceptualize formative assessment as an ongoing process of monitoring learners' progresses of knowledge construction. Randall, Cheong, and Engelhard (2011), 103, study the effects on the students' performance with or without identified disabilities as result of modifying the resource guide and calculator used by the Georgia Criterion Referenced Competency Test. Chang, Plake, Kramer, and Lien (2011), id 104, identify guessing behaviors and test-taking effort by means of detection indices, which are identified as equations. Based on such indices, unique response time patterns and guessing patterns are identified for six ability groups. Frey and Seitz (2011), id 105, examine the usefulness of multidimensional adaptive testing for the assessment of student literacy in the Program for International Student Assessment.

On the other hand, Bolt and Newton (2011), id 106, develop the measurement and control of extreme response style to the analysis of rating data from multiple scales. They claim that the current strategy is able to accommodate conditions in which the substantive traits across scales correlate. Barker-Plummer, Cox, and Dale (2011), id 107, analyze the assessment made by the Grade Grinder tool for the translations of natural language to sentences into first-order logic. The purpose is to find out a wide range of misunderstandings and confusions that students struggle with. During one year more, they continue the work and publish an enhanced version of the approach (Barker-Plummer, Dale, & Cox, 2012), id 124. In another vein, Seo, Kang, Drummond, and Kim (2011), id

Table 21
Main traits of the EDM approach profile that depicts assessment approaches published in 2011.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
101	Rad et al. (2011)	Machine learning	Descriptive	Clustering	*M*: IBL, simple multi-criteria decision making	*A*: analytic hierarchy process, k-means
102	Hsu et al. (2011)	Statistic	Descriptive	Correlation analysis	*M*: support vector Machine *T*: latent semantic analysis	*A*: drawing individual, collective cognition circles *E*: descriptive statistical
103	Randall et al. (2011)	Statistic, probability	Predictive	Classification	*M*: IRT *T*: Rasch model	*F*: hierarchical generalized linear model, Rasch model many-facet
104	Chang et al. (2011)	Statistic, probability	Predictive	Classification	*M*: IRT	*E*: guessing behaviors index, individual guessing behaviors index, test-taking effort index, individual test-taking effort index
105	Frey and Seitz (2011)	Statistic, probability	Predictive	Classification	*M*: IRT	*F*: three parameter logistic test model, multidimensional three-parameter logistic test model
106	Bolt and Newton (2011)	Statistic, probability	Predictive	Classification	*M*: IRT *T*: correlation analysis	*F*: Akaike information criterion, Bayesian information criterion, multidimensional nominal response model
107	Barker-Plummer et al. (2011)	Statistic	Descriptive	Clustering	*T*: frequencies	*E*: statistical
108	Seo et al. (2011)	Machine learning, DP	Predictive	Classification	*M*: HMM, decision tree	*A*: linear-chain conditional random field, J48
109	Kinnebrew and Biswas (2011)	DP	Descriptive	Sequential pattern	*M*: HMM	*A*: Baum-Welch, Pex-SPAM
110	Trivedi et al. (2011)	Machine learning	Descriptive, predictive	Clustering, classification	*M*: spectral clustering *T*: linear regression	*A*: spectral clustering, bootstrap aggregating ensamble
111	Ignatov et al. (2011)	Algebra	Descriptive	Clustering	*T*: formal concept analysis	*F*: lattice-based taxonomies
112	Romashkin et al. (2011)	Algebra	Descriptive	Clustering	*T*: formal concept analysis	*F*: lattice-based taxonomies
113	Worsley and Blikstein (2011)	Probability	Descriptive	Clustering	*M*: Bayes theorem	*A*: EM
114	Wauters, Desmet, and van den Noortgate (2011b)	Probability	Predictive	Classification	*M*: IRT	*A*: Elo rating
115	Feng et al. (2011)	Probability	Predictive	Regression	*T*: linear regression	*A*: stepwise linear regression

108, characterize successful versus unsuccessful question and answer type discussions to assess student learning in online discussion. So, they classify patterns of interactions using a state transition model and identify such a type of discussion. Kinnebrew and Biswas (2011), id 109, use HMM and sequence mining techniques to model learning behaviors from their interaction traces with the aim at comparing groups of students.

As well as Trivedi, Pardos, Sárközy, and Heffernan (2011), id 110, they predict student performance along two stages: firstly, clusters of students are produced by spectral clustering method; latterly, prediction of the student performance is fulfilled by a bootstrap aggregating ensamble algorithm. Ignatov, Mamedova, Romashkin, and Shamshurin (2011), id 111, build lattice-based taxonomies to depict the structure of the assessment data to identify the most stable student groups. They, Romashkin, Ignatov, and Kolotova (2011) id 112, also apply lattice-based taxonomies to analyze university applications. Worsley and Blikstein (2011), id 113, propose using student speech and drawings to decipher

Table 22
Main traits of the EDM approach profile that depicts assessment approaches published in 2012.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
116	Sen et al. (2012)	Machine learning, probability	Predictive	Classification, regression	*M*: decision tree, support vector machines sensitive analysis, neural networks, *T*: logistic regression	*A*: C5, multi-layer perceptron, nonlinear kernel functions, multiple logistic regression
117	Kobrin et al. (2012)	Statistic, probability	Predictive	Regression	*T*: correlation analysis, multiple regression	*E*: descriptive statistic, multiple regression
118	Mislevy et al. (2012)	Machine learning,	Descriptive	Clustering	*M*: IRT, evidence models theory	*A*: natural language processing
119	Gobert et al. (2012)	Machine learning, probability	Predictive	Classification	*M*: Bayes theorem, HMM	*A*: BKT
120	Kim et al. (2012)	Machine learning,	Descriptive	Clustering	*M*: latent Dirichlet allocation	*A*: latent Dirichlet allocation
121	Kinnebrew and Biswas (2012)	Machine learning,	Descriptive	Sequential pattern	*M*: sequence mining	*A*: differential sequence mining, linear segmentation
122	Sudol et al. (2012)	DP	Predictive	Classification	*M*: HMM	*E*: maximum likelihood estimate
123	López et al. (2012)	Machine learning, probability	Descriptive, predictive	Clustering, classification	*M*: decision tree, rules induction, Bayes theorem neural networks	*A*: DINB, NaiveBayesSimple, J48, multilayer perceptron
124	Barker-Plummer et al. (2012)	Statistic	Descriptive	Clustering	*T*: frequencies	*E*: statistical

meaningful markers of expertise in an automated and natural fashion. Wauters, Desmet, and van den Noortgate (2011b), id 114, focus on promptly estimating the learner's proficiency level by weight adaptation in the Elo rating system. Feng, Heffernan, Pardos, and Heffernan (2011), id 115, apply dynamic testing for assessment in ITS.

3.4.3. Assessment approaches published in 2012

During 2012 only nine mature approaches are published; thus, in this subsection a profile of they is outlined, without including the one that was given in the previous subsection. Firstly, the Table 22 provides their EDM approach profile, and later on, a synopsis of the approaches is stated as follows.

The account starts with Sen, Uçar, and Denle (2012), id 116, who identify the factors that lead students to success or failure of placement tests. So, they develop models to predict secondary education placement test results, and use sensitivity analysis on those prediction models to identify the most important predictors. Kobrin, Kim, and Sackett (2012), id 117, study the merits and pitfalls of standardized tests with questions regarding their item characteristics. They focus on investigating the relationship between item characteristics and the item's ability to predict college outcomes. Mislevy, Behrens, Dicerbo, and Levy (2012), id 118, define two viewpoints in educational assessment: one pursues structuring situations to evoke particular kinds of evidence, the other aims at discovering meaningful patterns in available data. Gobert, Sao Pedro, Baker, Toto, and Montalvo (2012), id 119, assess students' inquiry skills as they engage in an inquiry using science Microworlds. They also apply evidence-centered design framework to make inferences about student inquiry skills using models developed through a combination of text replay tagging and EDM.

In another vein, Kim, Shaw, Xu, and Adarsh (2012), id 120, implement TeamAnalytics, an instructional tool that facilitates the analyses of the student collaboration process by creating dynamic summaries of team member contributions over time. Kinnebrew and Biswas (2012), id 121, analyze transformation of action sequences using action features by contextualizing the sequence mining with information on the student's task performance and learning activities. Sudol, Rivers, and Harris (2012), id 122, tailor a metric to measure the probabilistic distance between an observed student solution and a correct solution. López, Luna, Romero, and Ventura (2012), 123, design a classification approach via clustering task to predict the final marks in a course on the basis of forum data. They conclude the EM clustering algorithm yields results similar to those of the best classification algorithms, especially when using only a group of selected attributes.

3.5. Student support and feedback

During the interaction between learner and system, the *student support* given by the computerized educational system is relevant to enhance the learners' performance and achievements, or to correct their misconceptions, bugs, and faults. Both kinds of consequences are respectively called *pre-emptive* and *post-failure*. In addition, most of the educational systems should offer functional-

Table 23
Main traits of the EDM approach profile that depicts student support and feedback approaches published from 2010 to 2012.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
125	Hsieh and Wang (2010)	Machine learning	Descriptive	Association rules	*T*: formal concept analysis	*A*: Apriori, learning path generation preference-based, correlation-based *F*: concept lattice
126	D'Mello, Olney, and Person (2010)	Probability	Descriptive	Sequential pattern	*M*: Bayes theorem	*E*: likelihood metric
127	Gupta and Rosé (2010)	Probability	Descriptive	Clustering	*M*: IBL	*A*: k-means
128	Dominguez et al. (2010)	Machine learning, statistic	Descriptive	Clustering, association rules, numerical analysis	*M*: IBL, decision tree *T*: frequencies-averages	*A*: k-means, Apriori *E*: statistical
129	Lehman et al. (2010)	Statistic	Descriptive	Clustering	*T*: frequencies, variability	*E*: statistical
130	Southavilay, Yacef, and Calvo (2010)	Machine learning	Descriptive	Sequential pattern	*T*: performance sequence analysis	*A*: performance sequence analysis
131	Kim and Calvo (2010)	Statistic	Descriptive	Clustering	*T*: latent semantic analysis non-negative matrix factorization	*A*: majority class baseline, keyword spotting, latent semantic analysis- based categorical classification, non-negative matrix *E*: factorization-based categorical classification
132	Champaign and Cohen (2010)	Probability	Predictive	Classification	*T*: probabilistic	*A*: dealing with the annotations
133	Vialardi et al. (2011)	Machine learning,, probability	Predictive	Classification	*M*: decision tree, ensamble, Bayes theorem, distance-based learning	*A*: NaiveBayes, C4.5, KNN, boosting
134	Khoshneshin et al. (2011)	Probability	Descriptive	Clustering	*M*: latent Dirichlet allocation	*A*: latent Dirichlet allocation
135	Jin et al. (2011)	DP	Predictive	Categorization	*M*: HMM	*A*: Markov decision process *F*: condensed linkage graphs data model
136	Tsuruta et al. (2012)	DP probability	Predictive	Classification	*M*: decision tree HMM	*A*: Markov decision processes *F*: Bayesian belief network
137	Leong et al. (2012)	Statistic	Descriptive	Clustering	*T*: descriptive statistic	*E*: descriptive statistic

ities to track *students' feedback* with the purpose to express: suggestions, complains, requests, and evaluations. In order to identify 21 approaches published from 2010 up to 2012, this subsection is organized into three. Firstly, a series of 8 approaches in progress is presented. Afterwards, the EDM approach profile of the mature is illustrated in Table 23. Finally, a series of 13 mature approaches is outlined.

As regards 2010, four incipient approaches oriented to student support are introduced next: Cade and Olney (2010) use a word's topic derived from a topic model to predict its label in two coding schemes of differing grain sizes; Boyer et al. (2010) apply hierarchical HMM for predicting tutor dialogue acts within a corpus; Stamper, Barnes, and Croy (2010) shape a method for generating a Bayesian knowledge base from a corpus of student problem attempt data to automatically generate hints for new students; Vialardi et al. (2010) develop an enrollment recommender system to assist students in their decision making for tailoring the right learning path.

Concerning 2011, two approaches in progress devoted to student feedback are presented: Koprinska (2011) investigates the effect of the stream on the student evaluation about teaching and the course marks. Barracosa and Antunes (2011b) propose a methodology to mine teaching behaviors from surveys filled by students, making use of some domain knowledge.

Regarding 2012, two student support approaches in progress are reported as follows: Johnson and Zaiane (2012) train medical students by the exposition of a broad range of examples supported by customized feedback and hints driven by an adaptive reinforcement learning system, named Shufti; Surpatean, Smirnov, and Manie (2012) analyze academic profile representations and similarity functions, as well as demonstrate how to combine recommender systems based on different similarity functions to achieve superior master recommendations.

As well as the mature approaches published in 2010, eight are found, six of them oriented to student support and two for student feedback. A profile of both kinds of targets is given in this paragraph and the following two: Hsieh and Wang (2010), id 125, support students when they survey and try to choose the right learning materials collected from Internet. Thus, they develop an approach to discover candidate courses, tailor a learning path, and recommend learning objects. D'Mello et al. (2010), id 126, outline a methodology to examine the structure of tutorial dialogues between students and expert human tutors by mining frequently occurring sequences of dialogue moves generated during tutorial sessions. Gupta and Rosé (2010), id 127, present an exploration of search behavior, search success, and recommendations for support students that are based on a data-driven methodology.

Concerning Dominguez, Yacef, and Curran (2010), id 128, they present a system that generates hints for students who are completing programming exercises. The system analyses clusters of patterns that affect the learners' performance during their interaction with the system. The clusters shape the basis for providing hints in real time. Lehman, Cade, and Olney (2010), id 129, study the effect produced by off topic conversation held by student-expert tutor. They apply "dialogue move occurrence" and "linguistic inquiry and word count analysis" to determine the anatomy of off topic conversation. Southavilay, Yacef, and Calvo (2010), id 130, develop heuristics to extract the semantic nature of text changes during collaborative writing. These semantic changes are then used to identify writing activities in writing processes.

Regarding the mature approaches for student feedback, the first is presented by Kim and Calvo (2010), id 131, who apply category-based and dimension-based emotion prediction models. Both models are inferred from textual and quantitative students' responses to Unit of Study Evaluations questionnaires with the aim at providing a comprehensive understanding of the student experience. The second is stated by Champaign and Cohen (2010), id 132, who explore the use of student annotations that allow students to leave comments on learning objects they are interacting with. Later on, the annotations are intelligently shown to similar students, who could identify which of them are useful.

As for 2011, three mature approaches exclusively oriented to student support are summarized as follows: Vialardi et al. (2011), id 133, design the Student Performance Recommender System to support the enrollment process using the students' academic performance record. The system estimates the inherent difficulty of a given course and measures the competence of a student for a course based on the grades obtained in related. Khoshneshin, Basir, Srinivasan, Street, and Hand (2011), id 134, apply a topic model to analyze the temporal change in the spoken language of a science classroom based on a dataset of conversations between a teacher and students. Jin et al. (2011), id 135, try to automate the creation of hints from student problem-solving data. Therefore, they design a technique to represent, classify, and use programs written by novices as a base for automatic hint generation for programming cognitive tutors.

With respect to 2012, two mature approaches are found, one per target. Regarding student support, Tsuruta, Knauf, Dohi, Kawabe, and Sakurai (2012), id 136, aim at solving the complicated University's system of course offerings, registration rules, and prerequisite courses, which should be matched to students' dynamic learning needs and desires. So, they develop a system to evaluate and refine curricula to reach an optimum of learning success in terms of best possible accumulative grade point average. Leong, Lee, and Mak (2012), id 137, fulfill sentiment mining for analyzing short message service texts in teaching evaluation. Data preparation involves the reading, parsing, and categorization of the texts. An interestingness criterion selects the sentiment model from which the sentiments of the students towards the lecture are discerned.

3.6. Curriculum, domain knowledge, sequencing, and teachers support

Curriculum is an essential labor of academics and teachers to develop before delivering instruction to their pupils. They spend a lot of time and effort engaged in authoring, seeking, adapting, and sequencing content resources to deploy the *curriculum*. According to differentiated instruction paradigm, academics are involved in the customization of curriculum and teaching practices with the aim at facilitating learners the acquisition of domain knowledge. Furthermore, content represents the *domain knowledge* repositories and cognitive models of the knowledge components to be learned and skills to be trained. Both items curriculum and content are delivered to students by the *sequencing* schema. Sequencing sketches action courses, evaluate options, and decides the teaching-learning experiences to deliver. What is more, *teachers support* represents the services devoted to facilitate the ordinary work performed by academics, such as: monitoring students, content searching, collaboration, and teachers modeling.

In order to depict the 19 approaches published from 2010 up to 2012 this section is organized into three parts to provide the next subjects: a series of 6 approaches in progress, the EDM approach profile of 13 mature approaches by means of Table 24, a series of mature approaches. The approaches of both series, incipient and mature, are ordered in the following sequence: curriculum, domain knowledge, sequencing, and teachers support. Moreover, the approaches are chronologically stated in progressive year of publication.

Concerning the incipient approaches, the series embraces five devoted to domain knowledge and one to sequencing. A profile of them is introduced next: Hardof-Jaffe, Hershkovitz, Azran, and Nachmias (2010) study the types of online hierarchical structures

Table 24
Main traits of the EDM approach profile that depicts curriculum, domain knowledge, sequencing, and teachers support approaches published from 2010 to 2012.

ID	Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
138	Maull, Saldivar, and Sumner (2010a)	Machine learning, probability	Descriptive	Clustering	*M*: IBL,Bayes theorem	*A*: k-means, EM
139	Vuong et al. (2011)	Statistic	Descriptive	Functional dependency	*T*: dependency structure	*E*: overall graduation rate
140	Brunskill (2011)	Probability	Descriptive	Clustering	*M*: Bayes theorem	*A*: EM
141	Durand, LaPlante, and Kop (2011)	DP	Predictive	Classification	*M*: HMM	*A*: Markov decision processes
142	Su et al. (2011)	Machine learning	Descriptive	Clustering	*T*: frequencies, variability	*E*: statistical
143	Hershkovitz et al. (2011a)	Machine learning, statistic	Descriptive	Sequential pattern	*T*: hierarchical clustering	*A*: two-step clustering *E*: descriptive statistic
144	Chi et al. (2011a)	DP	Predictive	Classification	*M*: HMM	*A*: Markov decision process, four reinforcement learning based feature-selection
145	Wauters, Desmet, and van den Noortgate (2011a)	Probability	Predictive	Classification	*M*: latent class analysis	*A*: latent class analysis modeling, clustering with latent class analysis
146	Xu and Recker (2011)	Statistic, probability	Descriptive	Clustering	*M*: decision tree, ensamble, Bayes theorem, distance-based learning	*A*: NaiveBayes, C4.5, KNN, boosting
147	Barracosa and Antunes (2011a)	Machine learning	Descriptive, predictive	Classification, sequential pattern	*M*: context-free language	*A*: pushdown automata *F*: meta-patterns
148	Cai et al. (2011)	Statistic	Predictive	Classification	*T*: descriptive statistic	*E*: statistical
149	Gaudioso et al. (2012)	Machine learning, probability	Predictive	Classification	*M*: decision tree, Bayes theorem, rules induction	*A*: naiveBayes, J48, Jrip, PART
150	Scheihing, Aros, and Guerra (2012)	Machine learning	Descriptive	Clustering	*M*: latent class analysis, IBL	*A*: k-means

of content items presented to students in LMS; Možina et al. (2010) design an algorithm that enables teachers to conceptualize procedural knowledge; Xu and Recker (2010) examine how the teacher uses the authoring tool Instructional Architect and conducts a SNA to characterize teachers' networked relationships; Koedinger and Stamper (2010) test several cognitive models that measure student acquisition of knowledge components; Zapata-Gonzalez, Menendez, Prieto- Mendez, and Romero (2011) propose a hybrid recommendation method to assist users in personalized searches for learning objects in repositories. As for the sequencing incipient approach, it is carried out by Brunskill and Russell (2011), whose module automatically constructs adaptive pedagogical strategies for ITS.

In another vein, the series of four curriculum mature approaches starts with Maull, Saldivar, and Sumner (2010b), id 138, who model and discover patterns of how teachers use the tool. The results reveal: teachers are engaging in behavior that shows affinity for the use of interactive digital resources as well as social sharing behaviors. Vuong, Nixon, and Towle (2011), id 139, assert: large-scale assessment data can be analyzed to determine the dependency relationships between units in a curriculum. Brunskill (2011), id 140, determines prerequisite structure from noisy observations of student data by estimating the probability of the observations under different possible prerequisite structures. Durand et al. (2011), id 141, help teachers to design their learning activities by intelligent components. So, they build the Intelligent Learning Design Recommendation system to propose learning paths in a LMS. During the learning design phase, the system exploits learning styles and teaching styles to support teachers work.

On the other hand, a pair of domain knowledge mature approaches is highlighted next. The first is a personalized learning content adaptation mechanism tailored by Su, Tseng, Lin, and Chen (2011), id 142. It manages historical learners' requests besides intelligently and directly delivering proper personalized learning content by means of adaptation decision and content synthesis processes. Concerning the second, Hershkovitz, Azran, Hardof-Jaffe, and Nachmias (2011), id 143, examine the use of online repositories of content items in hundreds of courses. Furthermore, courses with a significant repository size are analyzed to reveal hierarchical structures and identify associations between these structures and course characteristics.

Concerning the sequencing mature approaches, a pair is presented next: Chi, VanLehn, Litman, and Jordan (2011), id 144, implement a reinforcement learning approach for inducing pedagogical strategies and empirical evaluations of the induced strategies. Based on the reinforcement learning induced strategies, the ITS effectiveness improves the students' learning gains. As well as Wauters, Desmet, and van den Noortgate (2011a), id 145, they achieve adaptive item sequencing through items with a known difficulty level. The level is calibrated by the IRT, in which the item difficulty is matched to the learner knowledge level.

Finally, a series of five teaching support mature approaches begins with: Xu and Recker (2011), id 146, who assist teachers in accessing a digital library service to seek educational content. So, they apply latent class analysis to group teacher users according to their online behaviors. They find clusters of teachers ranging from window shoppers, lukewarm users to the most dedicated users. Barracosa and Antunes (2011a), id 147, tailor a methodology for anticipating teachers' performance based on the analysis of pedagogical surveys. Such an approach combines sequential pattern and classification to identify patterns used to enrich source data. In this way, teachers are characterized and the accuracy of the classification models is improved. Cai, Jain, Chang, and Kim (2011), id 148, sketch a model of teacher mentoring behaviors within a social networking site. The approach depicts how teachers help others in discussion forums and learn a model that characterizes their behavior. This mentoring model is consistent with profile answers based on the observed teachers' behaviors.

As well as Gaudioso, Montero, and Hernandez-del-Olmo (2012), id 149, they support teachers to monitor, understand, and evaluate the students' activity, particularly when students face problems. So, they develop predictive models to assist teachers during the students' activity. Scheihing, Aros, and Guerra (2012), id 150, analyze the interactions between teachers from many schools who collaborate through the Kelluwen network deployed on Web 2.0

3.7. Tools

A relevant contribution of the EDM to the DM field is the design, development, and testing of tools oriented to perform specific labors in the whole context of KDD. As a consequence of the diversity of functions performed in educational settings, diverse EDM tools are found during the analysis of the EDM works. However, three kinds of EDM tools are defined to gather those oriented to offer similar support. Thus, a profile of the three kinds of EDM tools is given in this Section. Such kinds are sampled by their respective instances in Table 25, where 18 tools published from 2010 up to 2012 are stated.

The first kind gathers four tools oriented to extraction, learning support, and feature engineering. Some of them facilitate *extraction* processes devoted to search, represent, and store raw data from educational systems into a suitable format to be mined. A different target is *learning support*, whose purpose is to facilitate knowledge acquisition and solve problems. *Feature engineering* is another target tackled by some tools with the aim at analyzing and choosing valuable attributes to be mined.

The second kind embraces six *visualization* tools to support the mining process, the analysis of results, and the interpretation of outcomes. Some of the tools illustrated in Table 25 facilitate monitoring learner activity, others make easy the design of curricula, and several discover patterns.

The third kind of tools contains eight instances devoted to *analysis support*. They deploy functionalities oriented to: evaluate the behavior and performance of the students during their interaction with the CBES, developing cognitive skills, helping to solve problems, and more purposes.

4. Discussion

Based on the sample of EDM works introduced in Section 3, a discussion about the results of the survey is outlined in this Section. In a sense, this overview is a kind of apology for encouraging the application of DM to the educational arena in order to discover useful knowledge. Thus, besides the motivations given by the prior sections, the first additional stimulus is the exposition of an anal-

Table 25
Main traits of the EDM tools published from 2010 to 2012.

ID	Author of the EDM tool	Type	Name	Purpose
1	Krüger, Merceron, and Wolf (2010)	1. Extraction	ExtractAndMap	Depicts and deploys functionalities concerning data extraction from LMS
2	Pedraza-Pérez, Romero, and Ventura (2011)	1. Extraction	Java desktop Moodle mining	Offers a wizard to facilitate the extraction of log data and the execution of DM processes
3	Mostafavi, Barnes, and Croy (2011)	1. Learning support	Logic Question Generator,	Generates proof problems that support and satisfy the conceptual requirements of the course instructor of deductive logic
4	Rodrigo, Baker, McLaren, Jayme, and Dy (2012)	1 Feature engineering	Workbench	Seeks and suggest appropriate features of educational settings like ITS
5	Johnson and Barnes (2010)	2. Visualization	InfoVis	Monitors students at learning to facilitate the supervision of the tutor
6	Macfadyen and Sorenson (2010)	2. Visualization	Learner Interaction Monitoring System	Captures data demonstrating learner online engagement with course materials
7	Maull, Saldivar, and Sumner (2010a)	2. Visualization	Curriculum Customization Service	Supports online curriculum planning and observe the behavior of teachers elicited during the curriculum planning
8	Rabbany, K., M., and O. R. (2011)	2. Visualization	Meerkat-ED	Tailors and visualizes snapshots of participants in the discussion forums, their interactions, and the tracking of the leader/peripheral students
9	García-Saiz and Zorrilla (2011)	2. Visualization	e-Learning Web Miner	Discovers students' behavior profiles and models about how they navigate and work in LMS
10	Johnson, Eagle, Joseph, and Barnes (2011)	2. Visualization	EDM Vis	Facilitates the visualization of information to explore, navigate, and understand learner data logs
11	Cohen and Nachmias (2011)	3. Analysis support	Web-log based	Evaluates pedagogical processes occurring in LMS settings and students' attitudes
12	García, Romero, Ventura, and de Castro (2011)	3. Analysis support	Continuous Improvement of e-Learning Courses Framework	Uncovers relationships discovered in student usage data through If–Then recommendation rules; as well as shares and score the rules previously obtained by instructors in similar courses with other instructors and experts in education
13	Fritz (2011)	3. Analysis support	Check My Activity	Support students to compare their own activity in Blackboard versus an anonymous summary of their course peers
14	Anjewierden, Gijlers, Saab, and De-Hoog (2011)	3. Analysis support	Brick	Explores patterns from action sequences derived from a simulation-based inquiry learning environment in which learners collaborate in dyads
15	Moreno, González, Estévez, and Popescu (2011)	3. Analysis support	SIENA	Achieves intelligent evaluation of knowledge building socially using conceptual maps with multimedia learning nodes
16	Dyckhoff, Zielke, Chatti, and Schroeder (2011)	3. Analysis support	eLAT	Enables teachers to explore and correlate content usage, user properties, user behavior, and assessment results through graphical indicators
17	Devine, Hossain, Harvey, and Baur (2011)	3. Analysis support	Data Miner for Outcomes based Education	Supports tutors analysis of learning results and performance records of their students. It uses supervised feature selection to produce learning patterns and interprets results to provide insights for course optimization.
18	Pechenizkiy, Trcka, Bra, and Toledo (2012)	3. Analysis support	CurriM	Analyzes student and education responsible perspectives on curriculum mining and shows the achievements of a project interested in developing curriculum

ysis about the functionalities and tools earlier presented, as well as a mention concerning their evolution. The second encouragement is the description of patterns that characterize EDM approaches as a result of the mining achieved by a DM application that was performed as part of our KDD method.

4.1. Analysis of the educational data mining functionalities

According to the arguments stated in Section 2.5.2, the sample of 240 EDM works was split into seven clusters, where six represent educational functionalities and one integrates EDM tools. A specific functionality reveals the essence of several EDM approaches, their purpose and application target. Thus, 222 EDM approaches were organized according to such a criterion to tailor a conceptual work area; where researchers, developers, and users, who hold similar interests, collaborate to spread a particular functionality.

Thus, in this subsection a brief analysis based on the individual works that compose a particular functionality or a set of tools is stated, and a highlight related to its evolution is also included. Therefore, during the reading of the analysis given by the following sub-subsections (e.g., 4.1.1,..., 4.1.7), the respective source (e.g., Sections 3.1 to 3.7) and the statistical information earlier provided (e.g., Tables 11–25) should be taken into account to facilitate the interpretation.

4.1.1. Analysis of student modeling

As for student modeling, it represents the kernel of the EDM labor. Excepting behavior and performance, all kinds of student traits, actions, and achievements are considered as part of this functionality that is simple called "student modeling". According to the analysis of the 43 EDM approaches identified in Section 3.1 the following traits are the most characterized: groups of activities, instruction and learning styles, resource usage, analysis and prediction of academic achievements, student success factors, concentration, postural patterns, students' mental states, domain knowledge, learning trajectories, knowledge tracing, and skills.

However, before the diversity of traits and the difficulty to depict them, the specialization of the student modeling functionality is demanded. Even though 66% of the approaches is mature and no incipient works were found in 2012, the total number is decreasing from 17 in 2010 up to 10 in 2012 as a consequence of the specialization effect. Thus, new mature targets could emerge as a result of the most claimed demands, such as: domain knowledge, skills, emotions, context, and cognition.

4.1.2. Analysis of student behavior modeling

In relation to student behavior modeling, it represents a mature target that challenges researchers to monitor, analyze, depict, simulate, and evaluate in real-time student behavior, as well as proactive and reactive modes. The analysis of the 48 EDM approaches stated in Section 3.2, where 66% is descriptive, unveils specific issues that are object of study and characterization such as: feature extraction, students contributions, persistence in online activity, careless attitude, gaming, metacognitive activity, user-system interaction, self-adaptation, collaborative activities, solving styles, as well as the prediction of student's ability, outcomes, understanding, behavior, task completion, and final marks.

The evolution of this functionality shows a slightly decreasing tendency of the quantity of approaches, whose annual average is 15. However, the demand for implementing real-time approaches for monitoring, supporting, and assessing student behavior should increase as much as online educational modalities (e.g., u-learning, educational networking...) spread.

4.1.3. Analysis of student performance modeling

With respect to student performance modeling, it is the "novelty" of the functionalities. As Section 3.3 reports, although its first approaches appeared in 2011, they have accumulated 46 works in just two years, a similar quantity to the prior functionalities. In consequence, 66% corresponds to incipient approaches and the remaining to matures, all of them oriented to deal with the following subjects: failure, success, students' response time, carelessness achievement, forgetting and relearning, time needed to solve a problem, sequential effect, preparation for future learning, knowledge mastered after a delay, disorientation, learning progression, response patterns, and learning achievements.

It is expected the performance functionality matures and their support be incorporated to CBES. Specially, the current demand is to provide timely students supervision and assessment with the aim at anticipating adjustments. This proactive policy is supported by the empirical evidence, because more than 80% corresponds to approaches based on predictive models.

4.1.4. Analysis of assessment

In another vein, assessment functionality represents the opportunity and responsibility to evaluate and control the efficacy, efficiency, quality, and degree of users' satisfaction of any kind of educational system, specially the CBES. As the sample embraces 45 EDM works, where 75% is mature, the interest in implementing this functionality is equivalent to the one held by the prior three. Practically, half the 45 approaches introduced in Section 3.4 is based on descriptive models and the remaining on predictive. The approaches are normally oriented to tackle the next issues: find recurrent failures, selection of examples, inquiry process, learned skills, discover relationships among responses, recruitment of aspirants, difficulty level of problems, student accuracy, learning activities, misunderstandings, confusions, merits and pitfalls of standardized tests.

Even though the sample shows a decreasing number of approaches published in 2012, assessment is an essential functionality to be permanently carried out. Moreover, new paradigms and modes of assessment are being considered such as: adaptive testing, self-assessment, dynamic prompting, and collaborative evaluation.

4.1.5. Analysis of student support and feedback

As regards student support and feedback, this functionality is the key to enhance the personalization and customization of CBES to meet students' demands. Furthermore, it is necessary to extend the scope, performance, and support provided by the whole educational system, including academic human resources. The series of 21 approaches stated in Section 3.5 unveils a halting occurrence of works, where nearly 40% is mature. Some of the study objects taken into account are the following: dialogue analysis, generation of hints, decision making, customized feedback, reinforcement, recommendations, opinion about teaching behaviors, advice content, student annotations, dealing with emotions, and stimulation of competences.

Concerning the functionality evolution, subjects such as the following are under development: recommender systems, adaptive support, text mining, web mining, analysis of tutorial and peer dialogues, and interaction through social networks.

4.1.6. Analysis of curriculum, domain knowledge, sequencing, and teachers support

The functionality that embraces curriculum, domain knowledge, sequencing, and teachers support represents heterogeneous tasks and components that CBES usually perform. The series of 19 approaches given in Section 3.6 uncovers a hesitant publication of works, where nearly 60% is incipient! Most of the approaches

provide the next kind of service: content authoring, knowledge description, teachers' collaboration to tailor curricula, personalized searching of educational content, user-tools interaction, curriculum analysis, scheduling of learning activities, design of hierarchical content structures, inducing pedagogical strategies, anticipation of teachers' performance, and teacher mentoring behaviors.

The evolution of this sort of functionality expects an increasing demand of the prior tasks and components, as well as the incursion in other targets, such as: development of adaptive pedagogical strategies, application of social networking to deliver education and spread collaboration, personalization of learning content, adaptive sequencing, and automation of content authoring.

4.1.7. Analysis of tools

As well as the set of 18 tools stated in Section 3.7, they are organized in three kinds of tools, where the first contains three sorts of tools. So, there are five types of tools (e.g., extraction, learning support, feature engineering, visualization, and analysis support) that denote the diversity of tasks, which are suitable to be supported. In spite of 60% of the tools being incipient and their publication is inconsistent, the claim for assigning devoted resources to build specialized EDM tools is growing.

Thus, the evolution of the development tools, as well as their application is oriented to: facilitate raw data extraction and transformation, simplify feature extraction, provide learning support to students, enhance the collaboration among peers of students as well as teachers, and introduce graphical and streaming interfaces to supervise students, data analysis, and knowledge discovery.

4.2. Discovery of patterns about educational data mining approaches

Based on the frequencies estimated for each value that instantiate a particular trait, a set of clusters was produced to identify the most usual value-instances for educational and DM traits. It means the resultant clusters unveil the *inner-relationships* between the values that instantiate a specific trait. In addition, seldom used traits to depict the EDM approaches were gathered into three clusters according to specific ranges of counting. In this way, the ground to support a proposal for a realistic pattern to depict EDM approaches was introduced through in Sections 2.4 and 2.5 through Tables 1–10.

However, the next issues are still pending to be solved: How is it possible to confirm such *inner-relationships*? What are the *inter-relationships* between the value-instances of the educational and DM traits? The responses to both questions will contribute to enforce the proposed *EDM approach pattern*.

The second answer reveals two specific patterns: one is the pattern for EDM approaches based on descriptive models, and the other is the pattern for EDM approaches based on predictive models, which are respectively called: *EDM descriptive approach pattern* and *EDM predictive approach pattern*.

4.2.1. Data mining application to unveil interrelationships between educational and data mining traits

With the aim of answering that pair of questions, a DM application was performed according to the next EDM approach profile: domain knowledge functionality; domain knowledge role; analysis role-type; evaluation module; monitor module-type; conventional system; personal system-name; machine learning discipline; descriptive model; clustering task; instance-based learning method; k-means algorithm. The implementation was made in Weka (Hall et al., 2009).

A sample of the outcome produced for the approach is shown in Table 26; where after 7 iterations a pair of clusters was generated. The cluster sum of squared errors was that 1444.2154. Although the binary vector holds 475 items, only those that represent the highest average of presence are shown. The first column identifies the numerical trait *id* and its name. The other three columns uncover the average estimated for the full data, cluster 0, and cluster 1, which respectively hold 222, 92, 130 records of the EDM approach profiles. In order to facilitate the interpretation, the name of the trait appears before its respective value-instances that reached the highest average.

Table 26
Outcome of the DM application to discover the interrelationships between educational and DM traits value-instances of 222 EDM approaches, which was deployed in Weka.

Educational and DM traits	Full data (222)	Cluster 0 (92)	Cluster 1 (130)
EDM functionalities			
2. Student behavior modeling	0.1937	0.1630	0.2154
3. Student performance modeling	0.2162	0.2609	0.1846
4. Assessment: general	0.2072	0.0761	0.3
5. Student modeling	0.2027	0.2391	0.1769
6. Student support and feedback	0.0946	0.1304	0.0692
7. Curriculum-domain knowledge-sequencing-teacher support	0.0856	0.1304	0.0538
Educational systems			
78 Conventional system	0.0901	0.1739	0.0308
91 Intelligent tutoring system	0.3964	0.1957	0.5385
93 Learning management system	0.0901	0.1304	0.0615
Educational systems-name			
107 Algebra	0.0901	0	0.1538
113 ASSISTments	0.0856	0.0326	0.1231
140 Moodle	0.0586	0.0761	0.0462
DM-Disciplines			
162 Machine learning	0.4054	0.5	0.3385
164 Probability	0.4550	0.2065	0.6308
165 Statistic	0.2117	0.337	0.1231
DM-Model			
167 Descriptive	0.4279	1	0.0231
168 Predictive	0.6126	0.0652	1
DM-Tasks			
169 Association rules	0.0721	0.1739	0
170 Classification	0.4595	0.0326	0.7615
173 Clustering	0.2928	0.6739	0.0231
177 Regression	0.1667	0.0326	0.2615
DM-Methods			
181 Bayes theorem	0.2162	0.0761	0.3154
194 Decision trees	0.1982	0.087	0.2769
208 Instances-based learning	0.0991	0.2065	0.0231
DM-Techniques			
242 Frequencies	0.045	0.1087	0
261 Logistic regression	0.0901	0	0.1538
DM-Algorithms			
325 Expectation maximization	0.0676	0.0652	0.0692
342 J48	0.0676	0.0326	0.0923
347 K-means	0.0856	0.1848	0.0154
373 NaiveBayes	0.0586	0.0109	0.0923
DM-Equations			
425 Descriptive statistic	0.027	0.0543	0.0077
451 Statistical	0.0946	0.1739	0.0385
DM-Frames			
461 Bayesian networks	0.0721	0	0.1231

4.2.2. Inner-relationships between value-instances of educational and data mining traits

With the purpose to respond to the first question and provide more evidence in favor of the EDM approach pattern, it is necessary to compare the averages shown in Table 26 against the counting presented in earlier Tables 1–10 stated in Sections 2.4 and 2.5. So, several matches between the average of specific value-instances and their respective counting estimated for them in their

respective "comparative table" are highlighted for the following traits[8]:

1. *EDM functionalities*: its value-instances 2 to 7 confirm their frequency unveiled in Table 10. So, the balanced proportion of the six functionalities that label EDM approaches (e.g., student modeling, assessment, curriculum...) is demonstrated.
2. *Educational systems*: the average of its value-instances 78, 91, and 93 is similar to the counting depicted in Table 8 for conventional, ITS, and LMS; therefore, it is confirmed the most common educational system are: ITS.
3. *Educational systems-name*: the average of its value-instances 107, 113, and 140 demonstrate that Algebra, ASSISTments, and Moodle are three of the most specific educational systems mined by EDM approaches as Table 9 asserts.
4. *Disciplines*: the average of its value-instances 162, 164, and 165 reinforce their prominence claimed in Table 1 for machine learning, probability, and statistic.
5. *Model*: the average of its value-instances 167 and 168 follow the same proportion as the one illustrated for descriptive and predictive in Fig. 2.
6. *Tasks*: the average of its value-instances 169, 170, 173, and 177 support their popularity manifested in Table 2 for association rules, classification, clustering, and regression.
7. *Methods*: the average of its value-instances 181, 194, and 208 guarantee that Bayes theorem, decision trees, and instances-based learning are the top-three methods unveiled in Table 3.
8. *Techniques*: the average of its value-instances 242 and 261 add evidence in favor of frequencies and logistic regression, the two techniques most common used according to Table 4.
9. *Algorithms*: the average of its value-instances 325, 342, 347, and 373 exclaim the most popular four algorithms are EM, J48, k-means, and Naive–Bayes, which have been identified in Table 5.
10. *Equations*: the average of its value-instances 425 and 451 coincide with Table 6 to assert statistical, including descriptive, equations are the most estimated by EDM approaches.
11. *Frames*: the average of its value-instance 461 confirms the most popular frame is Bayesian networks as Table 7 shows.

4.2.3. Interrelationships between value-instances of educational and data mining traits

In order to respond to the second question, an analysis of clusters 0 and 1 is demanded to discover the interrelationships between the value-instances of educational and DM traits. The right columns of Table 26 reveal the existence of a couple of "absolute" clusters; because the number 0 only contains EDM approaches whose model is descriptive; whereas, cluster 1 is exclusively made up of EDM approaches deployed as predictive models.

Therefore, two new patterns are yielded to characterize EDM approaches; the first is: both patterns, descriptive and predictive, are composed of the values that instantiate more often their respective EDM approaches than the ones which pertain to the other model. According to such kind of contrasting comparison, the following interrelationships between the value-instances of several traits are conjectured to tailor the respective pattern for descriptive and predictive EDM approaches:

Regarding EDM approaches characterized as descriptive models, they are suitable for functionalities such as: student performance modeling, student modeling, student support and feedback, and curriculum, domain knowledge, sequencing, and teacher support. They are used most often in conventional and LMS systems. Nearly the double of they, in comparison with predictive, discover knowledge from Moodle. Half these approaches is ground on the machine learning discipline, and one third lays on the statistic. Association rules are exclusively used by these approaches; whereas, clustering is nearly always applied by them too. Instances-based learning is their favorite method. Exclusively, they apply frequency techniques. What is more, they often deploy the k-means algorithm. The EM algorithm is similarly deployed by approaches based on descriptive as well as predictive models. Statistical equations, including descriptive, are preferably estimated by descriptive approaches; but, they rarely implement a frame.

As for EDM approaches based on predictive models, they are more in demand for student behavior modeling and assessment functionalities. Half is deployed in ITS. All the approaches that mine Algebra are depicted by predictive models, and the majority exploits ASSISTments. Approximately, two thirds is supported by the probability discipline. Classification and regression tasks are very often applied for predictive approaches. Bayes theorem and decision trees are preferable methods for predictive approaches. Only these approaches include logistic regression techniques. Besides, they frequently implement the J48 and Naïve–Bayes algorithms. The application of statistical equations by predictive approaches is unusual. In addition, they exclusively implement Bayesian networks.

5. Conclusions

The journey along the EDM arena is terminating. However, it is pertinent to visit two places before reaching the final point. The first highlights the main attributes and patterns found out from the sample of EDM works. The second develops an analysis of strengths, weaknesses, opportunities, and threats (SWOT) to describe the EDM arena status, as well as to provide a reference for future work. Finally, the following comment can be stated:

> "EDM is living its spring time and preparing for a hot summer season."

5.1. A snapshot of the survey of educational data mining works

Once the sample of EDM works has been presented, as well as several statistical and mining outcomes, a conceptual shape of the EDM field is sketched in this subsection. Such a viewpoint is made up of six subjects: one corresponds to the main EDM functionalities, the other concerns the EDM approach pattern, a pair is related to the EDM descriptive and predictive approach patterns, and two more are devoted to authors and institutions engaged in the development of the EDM arena.

The functionalities EDM (e.g., detailed by their specific Sections 3.1 to 3.6 and supported by statistical and mined outcomes given in Sections 2.4, 2.5, and 4.2) label the nature of the work being developed since 2010 up to date. Therefore, it is evident student modeling is the preferred target of research, development, and application, which is followed by assessment. Even though, the other three functionalities (e.g., student support..., curriculum..., tools) claim the focus and interest of the EDM community, as well as others that are identified in Section 1.

The EDM approach pattern (presented and ground in Section 4.2) highlights the preferences of the EDM community that lead the design, deployment, and operation of EDM approaches.

[8] In matches 3, 4, 6, 7, 8, 9, 10, and 11, the average of the full data estimated for the traits (i.e., second column in Table 26) is equivalent to the percentage computed for them in the respective comparative table (i.e., see the percentage column of the comparative table). The reason is: they apply a different denominator that corresponds to their respective sample size; in this case, the denominator is 222 (i.e., see the second column header of Table 26), which is different to the total counting stated at the bottom of the respective comparative table.

As has been stated, twelve traits make up the pattern, where four concern educational subject and eight the DM topic. The privileged values-instances of the traits shape the "favorite menu dishes". This means, the most common EDM approaches are oriented to student modeling and assessment functionalities, particularly those devoted to performance, behavior, learning, and domain knowledge. Such approaches mainly operate on ITS, LMS, and conventional systems. Particularly, they use Algebra, ASSISTments, and Moodle to mine their data. The DM profile of the majority of EDM approaches is supported by probability, machine learning, and static disciplines; where six of ten approaches are based on predictive models; whereas classification is the leader task that is followed by clustering; as well as, Bayes theorem and decision trees are the most used methods, which are complemented by logistic regression and frequencies techniques. Concerning the implementation, EDM approaches often turn to k-means, EM, J48, and Naive–Bayes algorithms; besides statistical and logistic regression equations, as well as, Bayesian networks, and its dynamical version, frame.

In another vein, the EDM descriptive approach pattern and EDM predictive approach pattern, earlier described in Section 4.2, unveil: how the EDM approaches probably look to each other based on their respective DM model. According to the privileged value-instances of the traits that make up both patterns, an additional remark is made: In some sense, descriptive and predictive patterns reflect the logistic behind the approaches developed during from 2010 to date, as well as being a reference worthy to be considered for future approaches.

On the other hand, the EDM community is composed of researchers and institutions. Given a segment of the references gathered in this survey, a ranking is estimated. It provides in descending order the authors and their number of EDM works as follows: (a) Ryan Baker: 15; (b) Cristobal Romero and Sebastián Ventura: 12; (c) Kenneth Koedinger: 11; (d) Joseph Beck: 7; (e) Kalina Yacef, Tiffany Barnes, John Stamper: 6; (f) Philip Pavlik, Jr., Neil T. Heffernan, Zachary Pardos, Sujith M. Gowda: 5.

As result of such a sample, it is desirable to award the support provided by some institutions involved in the EDM research, such as: Worcester Polytechnic Institute, USA; Carnegie Mellon University, USA; Cordoba University, Spain; University of Sydney, Australia; University of North Carolina at Charlotte, USA; University of Memphis, USA; Vanderbilt University, USA.

5.2. Analysis of the educational data mining field

The status of the EDM arena is briefly described by means of a SWOT analysis (Mengel, Sis, & Halloran, 2007). Besides of identifying strengths, weaknesses, opportunities, and threats about the EDM arena, these factors respectively represent: aspects to justify the use of EDM as part of the CBES scope, issues to overcome, profitable areas of development, and risks to be tackled by EDM community. The description of the four factors is given as follows.

As for the strengths, some of the most relevant are: the baseline is quite robust and mature, due to being supported by the DM and the educational systems fields, whose background was fully outlined in Sections 2.4 and 2.5. It represents a target of study and application for many disciplines that demand interdisciplinary and transdisciplinary viewpoints. It is perceived as a friendly environment among the EDM members, who share data logs, software, and findings. In addition, EDM specialized events and media are expanding quickly to spread the achievements and encourage new research projects.

Concerning the weaknesses, EDM is an incipient research field that is taken on few shoulders. Most of the recent works have been published in the proceedings of the International Conference on Educational Data Mining.[9] Specialized conferences and media in DM practically ignore the application of DM for education. Most of the EDM approaches represent the implementation of DM to explore educational subjects, instead of contributing to extend the DM field. Thus, many of the EDM researchers are in reality users of DM frameworks and tools. Because student modeling commands the focus of more than half the approaches, other targets are underdeveloped and others are still pending. According to the results analysis presented in Section 4.2, it is evident that most of the EDM approaches only apply a small portion of the huge repertory of DM items (e.g., disciplines, tasks, algorithms, equations, and frames) and ignore or underrate many other options.

The opportunities for EDM are tremendous! Education is a high priority of world society, which claims new paradigms to enhance the scope, quality, efficiency, and achievements of educational systems. Non-conventional educational approaches are welcome everywhere. Pedagogical paradigms demand student-centered education; as well as, personalized teaching is needed to meet individual, group, and community demands. Such a need is being tackled by AIWBES, where EDM represents an option to implement these properties. According to the fast evolution of computers, communications, internet, and heterogeneous platforms that facilitate the interaction man–machine anywhere/anytime, the sort of educational settings and data are growing exponentially and its diversification is unknown.

As regards the threats, they represent the constraints that avoid, delay, and obstruct the rational and formal development of EDM. Such barriers are mainly represented by the natural shortages that an incipient discipline has to confront. Therefore, EDM has to deal with: the lack of a particular theory to ground the EDM work essentials. What is more, a standard terminology, a common logistic, reliable frameworks, and open architectures are demanded to be proposed, accepted, and followed by the EDM community. Other issues are the lack of recognition and valorization of the contributions that EDM is able to provide for extending and enhancing the traditional achievements of educational systems.

Acknowledgments

The author gives testimony of the strength given by his Father, Brother Jesus, and Helper, as part of the research projects of World Outreach Light to the Nations Ministries (WOLNM). Moreover, a special mention is given to Leonor Adriana Cárdenas Robledo, who fulfilled a valueable task during the research, as well as Mr. Lawrence Whitehill Waterson, a native British English speaker who tunes the manuscript. Finally, this research holds a partial support from grants given by: CONACYT-SNI-36453, CONACYT 118862, CONACYT 118962-162727, IPN-SIP-20131093, IPN-COFAA-SIBE DEBEC/647-391/2013.

References

Akcapinar, G., Cosgun, E., & Altun, A. (2011). Prediction of perceived disorientation in online learning environment with random forest regression. In *Proceedings of the 4th international conference on educational data mining* (pp. 259–263).

Aleahmad, T., Aleven, V., & Kraut, R. (2010). Automatic rating of user-generated math solutions. In: *Proceedings of the 3rd international conference on educational data mining* (pp. 267–268).

Anaya, A. R., & Boticario, J. G. (2011a). Content-free collaborative learning modeling using data mining. *User Modeling and User-Adapted Interaction, 21*(1–2), 181–216.

Anaya, A. R., & Boticario, J. G. (2011b). Towards improvements on domain-independent measurements for collaborative assessment. In *Proceedings of the 4th international conference on educational data mining* (pp. 317–318).

Anjewierden, A., Gijlers, H., Saab, N., & De-Hoog, R. (2011). Brick: mining pedagogically interesting sequential patterns. In *Proceedings of the 4th international conference on educational data mining* (pp. 341–342).

[9] https://sites.google.com/a/iis.memphis.edu/edm-2013-conference/

Anjewierden, A., Kollöffel, B., & Hulshof, C. (2007). Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *Proceedings of the international workshop on applying data mining in e-Learning* (pp. 23–32).

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*(3), 383–398.

Arroyo, I., Mehranian, H., & Woolf, B. P. (2010). Effort-based tutoring: an empirical approach to intelligent tutoring. In *Proceedings of the 3rd international conference on educational data mining* (pp. 1–10).

Bachmann, M., Gobert, J., & Beck, J. (2010). Tracking students' inquiry paths through student transition analysis. In *Proceedings of the 3rd international conference on educational data mining* (pp. 269–270).

Baker, R. S. J. D., & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. In *Proceedings of the 3rd international conference on educational data mining* (pp. 11–20).

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future vision. *Journal of Educational Data Mining, 1*(1), 1–15.

Baker, R. S. J. D., Gowda, S. M., & Corbett, A. T. (2011). Automatically detecting a student's preparation for future learning: Help use is key. In *Proceedings of the 4th international conference on educational data mining* (pp. 179–188).

Baker, R. S. J. D., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G. W., Ocumpaugh, J., & Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In *Proceedings of the 5th international conference on educational data mining* (pp. 126–133).

Barker-Plummer, D., Cox, R., & Dale, R. (2011). Student translations of natural language into logic: The grade grinder corpus release 1.0. In *Proceedings of the 4th international conference on educational data mining* (pp. 51–60).

Barker-Plummer, D., Dale, R., & Cox, R. (2012). Using edit distance to analyse errors in a natural language to logic translation corpus. In *Proceedings of the 5th international conference on educational data mining* (pp. 134–141).

Barracosa, J., & Antunes, C. (2011a). Anticipating teachers' performance. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 77–82).

Barracosa, J., & Antunes, C. (2011b). Mining teaching behaviors from pedagogical surveys. In *Proceedings of the 4th international conference on educational data mining* (pp. 329–330).

Bayer, J., Bydzovská, H., Géryk, J., Obsıvac, T., & Popelínský, L. (2012). Predicting drop-out from social behaviour of students. In *Proceedings of the 5th international conference on educational data mining* (pp. 103–109).

Beheshti, B., Desmarais, M. C., & Naceur, R. (2012). Methods to find the number of latent skills. In *Proceedings of the 5th international conference on educational data mining* (pp. 81–86).

Bergner, Y., Dröschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data: machine-learning item response theory. In *Proceedings of the 5th international conference on educational data mining* (pp. 95–102).

Berkhin, P. (2006). Survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25–71). Heidelberg: Springer.

Bhattacharyya, D. K., & Hazarika, S. M. (2006). *Networks, data mining and artificial intelligence: trends and future directions*. New Delhi: Narosa Publishing House.

Bian, H. (2010). Clustering student learning activity data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 277–278).

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814–833.

Bouchet, F., Azevedo, R., Kinnebrew, J. S., & Biswas, G. (2012). Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning. In *Proceedings of the 5th international conference on educational data mining* (pp. 65–72).

Bousbia, N., Labat, J. M., Balla, A., & Rebai, I. (2010). Analyzing learning styles using behavioral indicators in web based learning environments. In *Proceedings of the 3rd international conference on educational data mining* (pp. 279–280).

Boyer, K. E., Phillips, R., Ha, E. Y., Wallis, M. D., Vouk, M. A., & Lester, J. C., (2010). A preliminary investigation of hierarchical hidden Markov models for tutorial planning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 285–286).

Brighton, H., & Mellish, C. (2002). Advances in instance selection for instance-based learning algorithms. *Journal of Data Mining and Knowledge Discovery, 6*, 153–172.

Brunskill, E. (2011). Estimating prerequisite structure from noisy data. In *Proceedings of the 4th international conference on educational data mining* (pp. 217–221).

Brunskill, E., & Russell, S. (2011). Partially observable sequential decision making for problem selection in an intelligent tutoring system. In *Proceedings of the 4th international conference on educational data mining* (pp. 327–328).

Buldua, A., & Üçgüna, K. (2010). Data mining application on students' data. *Procedia Social and Behavioral Sciences, 2*(2), 5251–5259.

Cade, W. L., & Olney, A. (2010). Using topic models to bridge coding schemes of differing granularity. In *Proceedings of the 3rd international conference on educational data mining* (pp. 281–282).

Cai, S., Jain, S., Chang, Y. H., & Kim, J. (2011). Towards identifying teacher topic interests and expertise within an online social networking site. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 97–102).

Campagni, R., Merlini, D., & Sprugnoli, R. (2012). Analyzing paths in a student database. In *Proceedings of the 5th international conference on educational data mining* (pp. 208–209).

Cetintas, S., Si, L., Xing, Y. P., Zhang, D., Park, J. Y., & Tzur, R. (2010). A joint probabilistic classification model of relevant and irrelevant sentences in Mathematical word problems. *Journal of Educational Data Mining, 2*(1), 83–101.

Champaign, J., & Cohen, R. (2010). An annotations approach to peer tutoring. In *Proceedings of the 3rd international conference on educational data mining* (pp. 231–240).

Chang, S. R., Plake, B. S., Kramer, G. A., & Lien, S. M. (2011). Development and application of detection indices for measuring guessing behaviors and test-taking effort in computerized adaptive testing. *Educational and Psychological Measurement, 71*(3), 437–459.

Chaturvedi, R. & Ezeife, C. I. (2012). Data mining techniques for design of its student models. In *Proceedings of the 5th international conference on educational data mining* (pp. 218–219).

Chau, M., Cheng, R., Kao, B., & Ng, J. (2006). Uncertain data mining: an example in clustering location data. In W. K. Ng, M. Kitsuregawa, J. Li, & K. Chang (Eds.), *Advances in knowledge discovery and data mining, lecture notes in computer science* (pp. 199–204). Heidelberg: Springer. Vol. 3918.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21*(1–2), 137–180.

Chi, M., Koedinger, K., Gordon, G., Jordan, P., & VanLehn, K. (2011). Instructional factors analysis: a cognitive model for multiple instructional interventions. In *Proceedings of the 4th international conference on educational data mining* (pp. 61–70).

Cobo, G., García, D., Santamaría, E., Morán, J. A., Melenchón, J., & Monzo, C. (2011). Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. In *Proceedings of the 4th international conference on educational data mining* (pp. 253–257).

Cohen, A., & Nachmias, R. (2011). What can instructors and policy makers learn about web-supported learning through web-usage mining. *The Internet and Higher Education, 14*(2), 67–76.

Crespo, P., & Antunes, C. (2012). Social networks analysis for quantifying students' performance in teamwork. In *Proceedings of the 5th international conference on educational data mining* (pp. 232–233).

Desmarais, M. C. (2011). Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In *Proceedings of the 4th international conference on educational data mining* (pp. 41–50).

Desmarais, M. C., & Pelczer, I. (2010). On the faithfulness of simulated student performance data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 21–30).

Devine, T., Hossain, M., Harvey, E. & Baur, A. (2011). Improving pedagogy by analyzing relevance and dependency of course learning outcomes. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 83–90).

Dominguez, A. K., Yacef, K., & Curran, J. R. (2010). Data mining for individualised hints in elearning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 91–100).

Durand, G., LaPlante, F. & Kop, R. (2011). A learning design recommendation system based on Markov decision processes. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 69–76).

Dyckhoff, A. L., Zielke, D., Chatti, M. A., & Schroeder, U. (2011). eLAT: an exploratory learning analytics tool for reflection and iterative improvement of technology enhanced learning. In *Proceedings of the 4th international conference on educational data mining* (pp. 355–356).

D'Mello, S., & Graesser, A. (2010). Mining bodily patterns of affective experience during learning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 31–40).

D'Mello, S., Olney, A., & Person, N. (2010). Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining, 2*(1), 1–37.

Eagle, M., Johnson, M., & Barnes, T. (2012). Interaction networks: generating high level hints based on network community clustering. In *Proceedings of the 5th international conference on educational data mining* (pp. 164–167).

Falakmasir, M. H., & Habibi, J. (2010). Using educational data mining methods to study the impact of virtual classroom in e-learning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 241–248).

Fancsali, S. (2011). Variable construction and causal modeling of online education messaging data: initial results. In *Proceedings of the 4th international conference on educational data mining* (pp. 331–332).

Fancsali, S. (2012). Variable construction and causal discovery for cognitive tutor log data: initial results. In *Proceedings of the 5th international conference on educational data mining* (pp. 237–238).

Feng, M., & Heffernan, N. (2010). Can we get better assessment from a tutoring system compared to traditional paper testing? Can we have our cake (better assessment) and eat it too (student learning during the test)? In *Proceedings of the 3rd international conference on educational data mining* (pp. 41–50).

Feng, M., Heffernan, N. T., Pardos, Z. A., & Heffernan, C. (2011). Comparison of traditional assessment with dynamic testing in a tutoring system. In *Proceedings of the 4th international conference on educational data mining* (pp. 295–300).

Fincham, J. M., Anderson, J. R., Betts, S., & Ferris, J. L. (2010). Using neural imaging and cognitive modeling to infer mental states while using an intelligent tutoring system. In *Proceedings of the 3rd international conference on educational data mining* (pp. 51–60).

Forsyth, C., Butler, H., Graesser, A. C., Halpern, D., Millis, K., Cai, Z., & Wood, J. (2010). Higher contributions correlate with higher learning gains. In *Proceedings of the 3rd international conference on educational data mining* (pp. 287–288).

Forsyth, C., Pavlik, Jr., P. I., Graesser, A. C., Cai, Z., Germany, M. L., Millis, K., Dolan, R. P., Butler, H., & Halpern, D. (2012). Learning gains for core concepts in a serious game on scientific reasoning. In *Proceedings of the 5th international conference on educational data mining* (pp. 172–175).

France, M. K., Finney, S. J., & Swerdzewski, P. (2010). Students' group and member attachment to their University: a construct validity study of the university attachment Scale. *Educational and Psychological Measurement, 70*(3), 440–458.

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the program for international student assessment. *Educational and Psychological Measurement, 71*(3), 503–522.

Fritz, J. (2011). Classroom walls that talk: using online course activity data of successful students to raise self-awareness of underperforming peers. *The Internet and Higher Education, 14*(2), 89–97.

Gandhi, P., & Aggarwal, V. (2010). Ensemble hybrid logit model. In *Proceedings of the KDD 2010 cup 2010: Workshop knowledge discovery in educational data* (pp. 33–50).

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education, 14*(2), 77–88.

García-Saiz, D. & Zorrilla, M. (2011). E-learning web miner: a data mining application to help instructors involved in virtual courses. In *Proceedings of the 4th international conference on educational data mining* (pp. 323–324).

García-Saiz, D., & Zorrilla, M. (2012). A promising classification method for predicting distance students' performance. In *Proceedings of the 5th international conference on educational data mining* (pp. 206–207).

Gaudioso, E., Montero, M., & Hernandez-del-Olmo, F. (2012). Supporting teachers in adaptive educational systems through predictive models: a proof of concept. *Expert Systems with Applications, 39*(1), 621–625.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Journal of Technometrics, 49*(3), 291–304.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 111–143.

Goguadze, G., Sosnovsky, S., Isotani, S., & Mclaren, B. M. (2011). Evaluating a bayesian student model of decimal misconceptions. In *Proceedings of the 4th international conference on educational data mining* (pp. 301–306).

Goldin, I. M., Koedinger, K. R., & Aleven, V. (2012). Learner differences in hint processing. In *Proceedings of the 5th international conference on educational data mining* (pp. 73–80).

Goldstein, A. B., Baker, R. S. J. D., & Heffernan, N. T. (2010). Pinpointing learning moments; a finer grain p(j) model. In *Proceedings of the 3rd international conference on educational data mining* (pp. 289–290).

Gong, Y., & Beck, J. E. (2011). Items, skills, and transfer models: which really matters for student modeling?. In *Proceedings of the 4th international conference on educational data mining* (pp. 81–90).

Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Using multiple Dirichlet distributions to improve parameter plausibility. In *Proceedings of the 3rd international conference on educational data mining* (pp. 61–70).

González-Brenes, J. P., & Mostow, J. (2010). Predicting task completion from rich but scarce data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 291–292).

González-Brenes, J. P., & Mostow, J. (2012). Dynamic cognitive tracing: towards unified discovery of student and cognitive models. In *Proceedings of the 5th international conference on educational data mining* (pp. 49–56).

González-Brenes, J. P., Duan, W., & Mostow, J. (2011). How to classify tutorial dialogue? comparing feature vectors vs. Sequences. In *Proceedings of the 4th international conference on educational data mining* (pp. 169–178).

Gowda, S. M., Baker, R. S. J. D., Pardos, Z. A., & Heffernan, N. (2011). The sum is greater than the parts: ensembling student knowledge models in ASSISTments. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 11–20).

Gowda, S. M., Rowe, J. P., Baker, R. S.J. D., Chi, M., & Koedinger, K. R. (2011). Improving models of slipping, guessing, and moment-by-moment learning with estimates of skill difficulty. In *Proceedings of the 4th international conference on educational data mining* (pp. 199–208).

Gupta, N. K., & Rosé, K. P. (2010). Understanding instructional support needs of emerging Internet users for Web-based information seeking. *Journal of Educational Data Mining, 2*(1), 38–82.

Guruler, H., Istanbullu, A., & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education, 55*(1).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations, 11*(1), 10–18.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Massachusetts: MIT Press.

Hardof-Jaffe, S., Hershkovitz, A., Azran, R., Nachmias, R. (2010). Hierarchical structures of content items in lms. In *Proceedings of the 3rd international conference on educational data mining* (pp. 293–294).

Hardoon, D. R., Shawe-Taylor, J., & Szedmak, S. (2004). Canonical correlation analysis: an overview with application to learning methods. *Journal of Neural Computation, 16*(12), 2639–2664.

He, Q., Veldkamp, B. P., & Westerhof, G. J. (2011). Computerized coding system for life narratives to assess students?' personality adaption. In *Proceedings of the 4th international conference on educational data mining* (pp. 325–326).

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*(1), 90–102.

Hershkovitz, A., & Nachmias, R. (2010). Is students' activity in lms persistent?. In *Proceedings of the 3rd international conference on educational data mining* (pp. 295–296).

Hershkovitz, A., & Nachmias, R. (2011). Online persistence in higher education web-supported courses. *The Internet and Higher Education, 14*(2), 98–106.

Hershkovitz, A., Azran, R., Hardof-Jaffe, S., & Nachmias, R. (2011). Types of online hierarchical repository structures. *The Internet and Higher Education, 14*(2), 107–112.

Hershkovitz, A., Baker, R. S. J. D., Gobert, J., & Wixon, M. (2011). Goal orientation and changes of carelessness over consecutive trials in science inquiry. In *Proceedings of the 4th international conference on educational data mining* (pp. 315–316).

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Oklahoma: StatSoft.

Holzhüter, M., Frosch-Wilke, D., & Klein, U. (2012). Exploiting learner models using data mining for e-learning: a rule based approach. In A. Peña-Ayala (Ed.), *Intelligent and adaptive educational- learning systems: achievements and trends, smart innovation, systems and technologies* (pp. 77–105). Heidelberg: Springer.

Hong, T. P., Lin, K. Y., & Wang, S. L. (2003). Fuzzy data mining for interesting generalized association rules. *Journal of Fuzzy Sets and systems, 138*(2), 255–269.

Howard, L., Johnson, J., & Neitzel, C. (2010). Examining learner control in a structured inquiry cycle using process mining. In *Proceedings of the 3rd international conference on educational data mining* (pp. 71–80).

Hsieh, T., & Wang, T. (2010). A mining-based approach on discovering courses pattern for constructing suitable learning path. *Expert Systems with Applications, 37*(6), 4156–4167.

Hsu, J., Chou, H., & Chang, H. (2011). EduMiner: using text mining for automatic formative assessment. *Expert Systems with Applications, 38*(4), 3431–3439.

Huei-Tse, H. (2011). A case study of online instructional collaborative discussion activities for problem-solving using situated scenarios: an examination of content and behavior cluster analysis. *Computers & Education, 56*(3), 712–719.

Ignatov, D., Mamedova, S., Romashkin, N., & Shamshurin, I. 2011. What can closed sets of students and their marks say? In *Proceedings of the 4th international conference on educational data mining* (pp. 223–228).

Inventado, P. S., Legaspi, R., Suarez, M., & Numao, M. (2011). Investigating the transitions between learning and non-learning activities as students learn online. In *Proceedings of the 4th international conference on educational data mining* (pp. 367–368).

Ivancevic, V., Celikovic, M., & Lukovic, I. (2011). Analyzing student spatial deployment in a computer laboratory. In *Proceedings of the 4th international conference on educational data mining* (pp. 265–270).

Jarusek, P., & Pelánek, R. (2011). Problem response theory and its application for tutoring. In *Proceedings of the 4th international conference on educational data mining* (pp. 371–372).

Jeong, H., Biswas, G., Johnson, J., & Howard, L. (2010). Analysis of productive learning behaviors in a structured inquiry cycle using hidden Markov models. In *Proceedings of the 3rd international conference on educational data mining* (pp. 81–90).

Jin, W., Lehmann, L., Johnson, M., Eagle, M., Mostafavi, B., Barnes, T., & Stamper, J. (2011). Towards Automatic Hint Generation for a Data-Driven Novice Programming Tutor. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 91–96).

Johnson, M. W., & Barnes, T. (2010). EDM visualization tool: watching students learn. In *Proceedings of the 3rd international conference on educational data mining* (pp. 297–298).

Johnson, M. W., Eagle, M. J., Joseph, L., & Barnes, T. (2011). The EDM Vis tool. In *Proceedings of the 4th international conference on educational data mining* (pp. 349–350).

Johnson, S., & Zaiane, O. (2012). Deciding on feedback polarity and timing. In *Proceedings of the 5th international conference on educational data mining* (pp. 220–221).

Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). Analyzing university data for determining student profiles and predicting performance. In *Proceedings of the 4th international conference on educational data mining* (pp. 347–348).

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms* (2nd ed.). New Jersey: IEEE Press.

Kardan, S., & Conati, C. (2011). A framework for capturing distinguishing user interaction behaviours in novel interfaces. In *Proceedings of the 4th international conference on educational data mining* (pp. 159–168).

Karegar, M., Isazadeh, A., Fartash, F., Saderi, T., & Navin, A. H. (2008). Data-mining by probability-based patterns. In *Proceedings of the 30th international conference on information technology interfaces* (pp. 353–360).

Kerr, D., & d Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1), 144–182.

Keshtkar, F., Morgan, B., & Graesser., A. (2012). Automated detection of mentors and players in an educational game. In *Proceedings of the 5th international conference on educational data mining* (pp. 212–213).

Khodeir, N., Wanas, N., Darwish, N., & Hegazy, N. (2010). Inferring the differential student model in a probabilistic domain using abduction inference in bayesian

networks. In *Proceedings of the 3rd international conference on educational data mining* (pp. 299–300).
Khoshneshin, M., Basir, M. A., Srinivasan, P., Street, N., & Hand, B. (2011). Analyzing the language evolution of a science classroom via a topic model. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 41–50).
Kim, J., Shaw, E., Xu, H., & Adarsh, G. V. (2012). Assisting instructional assessment of undergraduate collaborative wiki and SVN activities. In *Proceedings of the 5th international conference on educational data mining* (pp. 10–16).
Kim, S. M., & Calvo, R. A. (2010). Sentiment analysis in student experiences of learning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 111–120).
Kinnebrew, J. S., & Biswas, G. (2011). Comparative action sequence analysis with hidden Markov models and sequence mining. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 59–68).
Kinnebrew, J. S., & Biswas, G. (2012). Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In *Proceedings of the 5th international conference on educational data mining* (pp. 57–64).
Kobrin, J. L., Kim, J. K., & Sackett, P. R. (2012). Modeling the predictive validity of SAT mathematics items using item characteristics. *Educational and Psychological Measurement, 72*(1), 99–119.
Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction, 21*(1–2), 51–97.
Koedinger, K. R., & Stamper, J. C. (2010). A data driven approach to the discovery of better cognitive models. In *Proceedings of the 3rd international conference on educational data mining* (pp. 325–326).
Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated student model improvement. In *Proceedings of the 5th international conference on educational data mining* (pp. 17–24).
Koedinger, K. R., Pavlik, Jr., P. I., Stamper, J. C., Nixon, T., & Ritter, S. (2011). Avoiding problem selection thrashing with conjunctive knowledge tracing. In *Proceedings of the 4th international conference on educational data mining* (pp. 91–100).
Koprinska, I. (2011). Mining assessment and teaching evaluation data of regular and advanced stream students. In *Proceedings of the 4th international conference on educational data mining* (pp. 359–360).
Krüger, A., Merceron, A., & Wolf, B. (2010). A data model to ease analysis and mining of educational data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 131–140).
Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the graduate record examination for master's and doctoral programs: a meta-analytic investigation. *Educational and Psychological Measurement, 70*(2), 340–352.
Lee, J. I., & Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. In *Proceedings of the 5th international conference on educational data mining* (pp. 118–125).
Lehman, B., Cade, W., & Olney, A. (2010). Off topic conversation in expert tutoring: waste of time or learning opportunity?. In *Proceedings of the 3rd international conference on educational data mining* (pp. 101–110).
Lemmerich, F., Ifland, M., & Puppe, F. (2011). Identifying influence factors on students success by subgroup discovery. In *Proceedings of the 4th international conference on educational data mining* (pp. 345–346).
Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications, 39*(3), 2584–2589.
Levy, S. T., & Wilensky, U. (2011). Mining students' inquiry actions for understanding of complex systems. *Computers & Education, 56*(3), 556–573.
Li, N., Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2011). A machine learning approach for automatic student model discovery. In *Proceedings of the 4th international conference on educational data mining* (pp. 31–40).
Liu, K. & Xing, Y. (2010). A lightweight solution to the educational data mining challenge. In *Proceedings of the KDD 2010 cup 2010 workshop knowledge discovery in educational data* (pp. 76–82).
López, M. I., Luna, J. M., Romero, C., Ventura, S. (2012). Classification via clustering for predicting final marks starting from the student participation in forums. In *Proceedings of the 5th international conference on educational data mining* (pp. 148–151).
Luan, J. (2002). Data mining and its applications in higher education. *Journal of New Directions for Institutional Research, 113*, 17–36.
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.
Macfadyen, L. P., & Sorenson, P. (2010). Using LiMS (the learner interaction monitoring system) to track online learner engagement and evaluate course design. In *Proceedings of the 3rd international conference on educational data mining* (pp. 301–302).
Malmberg, J., Järvenoja, H., & Järvela, S. (2013). Patterns in elementary school student strategic actions in varying learning situations. *Instructional Science*, 1-22. In press, published on line January, 2013. http://link.springer.com/article/10.1007/s11251-012-9262-1#page-1.
Marquez-Vera, C., Romero, C., & Ventura, S. (2011). Predicting school failure using data mining. In *Proceedings of the 4th international conference on educational data mining* (pp. 271–275).
Martinez, R., Yacef, K., & Kay, J. (2012). Speaking (and touching) to learn: a method for mining the digital footprints of face-to-face collaboration. In *Proceedings of the 5th international conference on educational data mining* (pp. 234–235).
Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., & Kharrufa, A. (2011). Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Proceedings of the 4th international conference on educational data mining* (pp. 111–120).
Maull, K. E., Saldivar, M. G., & Sumner, T. (2010a). Observing online curriculum planning behavior of teachers. In *Proceedings of the 3rd international conference on educational data mining* (pp. 303–304).
Maull, K. E., Saldivar, M. G., & Sumner, T. (2010b). Online curriculum planning behavior of teachers. In *Proceedings of the 3rd international conference on educational data mining* (pp. 121–130).
McCarthy, P. M., & Boonthum-Denecke, Ch. (2011). *Applied natural language: identification, investigation and resolution*. Pensilvania: International Science Reference.
Mccuaig, J., & Baldwin, J. (2012). Identifying successful learners from interaction behaviour. In *Proceedings of the 5th international conference on educational data mining* (pp. 160–163).
Mengel, M., Sis, B., & Halloran, P. F. (2007). SWOT analysis of Banff: strengths, weaknesses, opportunities and threats of the international Banff consensus process and classification system for renal allograft pathology. *American Journal of Transplantation, 7*(10), 2221–2226.
Merceron, A. (2011). Investigating usage of resources in LMS with specific association rules. In *Proceedings of the 4th international conference on educational data mining* (pp. 361–362).
Merceron, A., Schwarzrock, S., Elkina, M., Pursian, A., Beuster, L., Fortenbacher, A., Kappe, L., & Wenzlaff, B. (2012). Learning paths in a non-personalizing e-learning environment. In *Proceedings of the 5th international conference on educational data mining* (pp. 228–229).
Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48.
Mitra, S., & Acharya, T. (2003). *Data mining: multimedia, soft computing, and bioinformatics*. New Jersey: John Wiley & Sons, Inc..
Molina, M. M., Luna, J. M., Romero, C., & Ventura, S. (2012). Meta-learning approach for automatic parameter tuning: a case of study with educational datasets. In *Proceedings of the 5th international conference on educational data mining* (pp. 180–183).
Montalvo, O., Baker, R. S. J. D., Sao-Pedro, M. A., Nakama, A., & Gobert, J. D. (2010). Identifying students' inquiry planning using machine learning. In *Proceedings of the 3rd international conference on educational data mining* (pp. 141–150).
Moreno, L., González, C., Estévez, R., & Popescu, B. (2011). Intelligent evaluation of social knowledge building using conceptual maps with MLN. In *Proceedings of the 4th international conference on educational data mining* (pp. 343–344).
Mostafavi, B., Barnes, T., & Croy, M. (2011). Automatic generation of proof problems in deductive logic. In *Proceedings of the 4th international conference on educational data mining* (pp. 289–294).
Mostow, J., González-Brenes, J., & Tan, B. H. (2011). Learning classifiers from a relational database of tutor logs. In *Proceedings of the 4th international conference on educational data mining* (pp. 149–158).
Mostow, J., Xu, Y., & Munna, M. (2011). Desperately seeking subscripts: towards automated model parameterization. In *Proceedings of the 4th international conference on educational data mining* (pp. 283–288).
Možina, M., Guid, M., Sadikov, A., Groznik, V., Krivec, J., & Bratko, I. (2010). Conceptualizing procedural knowledge targeted at students with different skill levels. In *Proceedings of the 3rd international conference on educational data mining* (pp. 309–310).
Muldner, K., Burleson, W., Van de Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User Modeling and User-Adapted Interaction, 21*(1–2), 99–135.
Nandeshwar, A., Menzies, T., & Nelson, A. (2013). Learning patterns of university student retention. *Expert Systems with Applications, 38*(12), 14984–14996.
Narli, S., Özgen, K., & Alkan, H. (2011). In the context of multiple intelligences theory, intelligent data analysis of learning styles was based on rough set theory. *Learning and Individual Differences, 21*(5), 613–618.
Nooraei, B., Pardos, Z. A., Heffernan, N., T., & Baker, R. S. J. D. (2011). Less is more: improving the speed and prediction power of knowledge tracing by using less data. In *Proceedings of the 4th international conference on educational data mining* (pp. 101–109).
Nugent, R., Dean, N., & Ayers, E. (2010). Skill set profile clustering: the empty k-means algorithm with automatic specification of starting cluster centers. In *Proceedings of the 3rd international conference on educational data mining* (pp. 151–160).
Nwaigwe, A. F., & Koedinger, K. R. (2011). The simple location heuristic is better at predicting students' changes in error rate over time compared to the simple temporal heuristic. In *Proceedings of the 4th international conference on educational data mining* (pp. 71–80).
Pardos, Z. A., & Heffernan, N. T. (2010a). Navigating the parameter space of bayesian knowledge tracing models: visualizations of the convergence of the expectation maximization algorithm. In *Proceedings of the 3rd international conference on educational data mining* (pp. 161–170).
Pardos, Z. A., & Heffernan, N. T. (2010b). Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. In *Proceedings of the KDD 2010 cup 2010 workshop knowledge discovery in educational data* (pp. 24–35).
Pardos, Z. A., Wang, Q. Y., & Trivedi, S. (2012). The real world significance of performance prediction. In *Proceedings of the 5th international conference on educational data mining* (pp. 192–195).
Pardos, Z. A., Gowda, S. M., Baker, R. S. J. D., & Heffernan, N. T. (2011). Ensembling predictions of student post-test scores for an intelligent tutoring system. In

Proceedings of the 4th international conference on educational data mining (pp. 189–198).
Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Educational and Psychological Measurement, 72*(1), 44–51.
Pavlik, Jr., P. I. (2010). Data reduction methods applied to understanding complex learning hypotheses. In *Proceedings of the 3rd international conference on educational data mining* (pp. 311–312).
Pavlik Jr., P. I., & Wu, S. (2011). A dynamical system model of microgenetic changes in performance, efficacy, strategy use and value during vocabulary learning. In *Proceedings of the 4th international conference on educational data mining* (pp. 277–282).
Pechenizkiy, M., Trcka, N., Bra, P. D., & Toledo, P. (2012). CurriM: curriculum mining. In *Proceedings of the 5th international conference on educational data mining* (pp. 216–217).
Peckham, T., & McCalla, G. (2012). Mining student behavior patterns in reading comprehension tasks. In *Proceedings of the 5th international conference on educational data mining* (pp. 87–94).
Pedraza-Pérez, R., Romero, C., & Ventura, S. (2011). A java desktop tool for mining moodle data. In *Proceedings of the 4th international conference on educational data mining* (pp. 319–320).
Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making, 7*(4), 639–682.
Peña-Ayala, A., Domínguez, R., & Medel, J. (2009). Educational data mining: a sample of review and study case. *World Journal of Educational Technology, 2*, 118–139.
Peña-Ayala, A. (Ed.). (2013). Educational data mining: applications and trends, studies in computational intelligence. Heidelberg: Springer Verlang. http://www.springer.com/series/7092.
Perez-Mendoza, R., Rubens, N., & Okamoto, T. (2010). Hierarchical aggregation prediction method. In *Proceedings of the KDD 2010 cup 2010 workshop knowledge discovery in educational data* (pp. 62–75).
Ping-Feng, P., Yi-Jia, L., & Yu-Min, W. b. (2010). Analyzing academic achievement of junior high school students by an improved rough set model. *Computers & Education, 54*(4), 889–900.
Psotka, J., Massey, L. D., & Mutter, S. A. (Eds.). (1989). *Intelligent tutoring systems: lessons learned*. New Jersey: Lawrence Erlbaum Associates, Inc..
Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., & Heffernan, N. T. (2011). Does time matter? modeling the effect of time in bayesian knowledge tracing. In *Proceedings of the 4th international conference on educational data mining* (pp. 139–148).
Rabbany, R., K., Takaffoli, M. & Zaiane, O. R. (2011). Analyzing participation of students in online courses using social network analysis techniques. In *Proceedings of the 4th international conference on educational data mining* (pp. 21–30).
Rad, A., Naderi, B., & Soltani, M. (2011). Clustering and ranking university majors using data mining and AHP algorithms. A case study in Iran. *Expert Systems with Applications, 38*(1), 755–763.
Rafferty, A. N., Lamar, M. M., & Griffiths, T. L. (2012). Inferring learners' knowledge from observed actions. In *Proceedings of the 5th international conference on educational data mining* (pp. 226–227).
Rai, D., & Beck, J. E. (2010). Analysis of a causal modeling approach: a case study with an educational intervention. In *Proceedings of the 3rd international conference on educational data mining* (pp. 313–314).
Rai, D., & Beck, J. E. (2011). Exploring user data from a game-like math tutor: a case study in causal modeling. In *Proceedings of the 4th international conference on educational data mining* (pp. 307–311).
Rajibussalim. (2010). Mining students' interaction data from a system that support learning by reflection. In *Proceedings of the 3rd international conference on educational data mining* (pp. 249–256).
Randall, J., Cheong, Y. F., & Engelhard, G. Jr., (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement, 71*(1), 129–147.
Rau, M. A., & Pardos, Z. A. (2012). Interleaved practice with multiple representations: analyses with knowledge tracing based techniques. In *Proceedings of the 5th international conference on educational data mining* (pp. 168–171).
Rau, M. A., & Scheines, R. (2012). Searching for variables and models to investigate mediators of learning from multiple representations. In *Proceedings of the 5th international conference on educational data mining* (pp. 110–117).
Rebak, D. R., Blackmon, B., & Humphreys, A. R. (2000). The virtual university: developing a dynamic learning management portal. *Journal of IEEE Concurrency, 8*(3), 3–5.
Rodrigo, M. M. T., Baker, R. S. J. D., McLaren, B., Jayme, A., & Dy, T. T. (2012). Development of a workbench to address the educational data mining bottleneck. In *Proceedings of the 5th international conference on educational data mining* (pp. 152–155).
Romashkin, N., Ignatov, D., & Kolotova, E. (2011). How university entrants are choosing their department? mining of university admission process with FCA taxonomies. In *Proceedings of the 4th international conference on educational data mining* (pp. 229–233).
Romero, C., & Ventura, S. (Eds.). (2006). *Data mining in E-learning. Advances in Management Information*. Wessex: Wit Press. vol. 4.
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., Romero, J. R., Luna, J. M., & Ventura, S. (2010). Mining rare association rules from e-learning data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 171–180).
Romero, C., Ventura, S., Pechenizkiy, M., & Ryan, S. J. d. (Eds.). (2011). *Handbook of educational data mining, data mining and knowledge discovery series*. Florida: Chapman & Hall/CRC.
Romero, C., Ventura, S., Vasilyeva, E., & Pechenizkiy, M. (2010). Class association rules mining from students' test data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 317–318).
Rupp, A. A., Sweet, S. J., & Choi, Y. (2010). Modeling learning trajectories with epistemic network analysis: a simulation-based investigation of a novel analytic method for epistemic games. In *Proceedings of the 3rd international conference on educational data mining* (pp. 319–320).
Rupp, A. A., Levy, R., Dicerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., et al. (2012). Putting ECD into practice: the interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4*(1), 49–110.
Rus, V., Moldovan, C., Niraula, N., & Graesser, A. C. (2012). Automated discovery of speech act categories in educational games. In *Proceedings of the 5th international conference on educational data mining* (pp. 25–32).
Sabourin, J. L., Mott, B. W., & Lester, J. C. (2012). Early prediction of student self-regulation strategies by combining multiple models. In *Proceedings of the 5th international conference on educational data mining* (pp. 156–159).
Sao-Pedro, M. A., Baker, R. S. J. D., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using text replay tagging to produce detectors of systematic experimentation behavior patterns. In *Proceedings of the 3rd international conference on educational data mining* (pp. 181–190).
Scheihing, E., Aros, C., & Guerra, D. (2012). Analyzing the behavior of a teacher network in a web 2.0 environment. In *Proceedings of the 5th international conference on educational data mining* (pp. 210–211).
Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior, 28*(4), 1321–1331.
Sen, B., Uçar, E., & Denle, D. (2012). Predicting and analyzing secondary education placement-test scores: a data mining approach. *Expert Systems with Applications, 39*(10), 9468–9476.
Seo, S. W., Kang, J. H., Drummond, J., & Kim, J. (2011). A state transition model for student online discussions. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 51–58).
Shanabrook, D. H., Cooper, D. G., Woolf, B. P., & Arroyo, I. (2010). Identifying high-level student behavior using sequence-based motif discovery. In *Proceedings of the 3rd international conference on educational data mining* (pp. 191–200).
Shen, Y., Chen, Q., Fang, M., Yang, Q., Wu, T., Zheng, L., & Cai, Z. (2010). Predicting student performance: a solution for the KDD. In *Proceedings of the KDD 2010 cup 2010 workshop: Knowledge discovery in educational data* (pp. 24–32).
Shih, B., Koedinger, K. R., & Scheines, R. (2010). Unsupervised discovery of student learning tactics. In *Proceedings of the 3rd international conference on educational data mining* (pp. 201–210).
Shu-Hsien, L., Pei-Hui, C., & Pei-Yuan, H. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303–11311.
Sohn, S. Y., & Ju, Y. H. (2010). Conjoint analysis for recruiting high quality students for college education. *Expert Systems with Applications, 37*(5), 3777–3783.
Songmuang, P., & Ueno, M. (2010). Multiple test forms construction based on bees algorithm. In *Proceedings of the 3rd international conference on educational data mining* (pp. 321–322).
Soundranayagam, H., & Yacef, K. (2010). Can order of access to learning resources predict success?. In *Proceedings of the 3rd international conference on educational data mining* (pp. 323–324).
Southavilay, V., & Yacef, K., & Calvo, R. A. (2010). Process mining to support students' collaborative writing. In *Proceedings of the 3rd international conference on educational data mining* (pp. 257–266).
Sparks, R. L., Patton, J., & Ganschow, L. (2012). Profiles of more and less successful L2 learners: a cluster analysis study. *Learning and Individual Differences, 22*(4), 463–472.
Srinivas, S., Hamby, S., Lofthus, R., Caruthers, E., Barett, J., & Ells, E. (2011). From data to actionable knowledge: a collaborative effort with educators. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 109–114).
Stamper, J. C., Barnes, T., & Croy, M. (2010). Using a Bayesian knowledge base for hint selection on domain specific problems. In *Proceedings of the 3rd international conference on educational data mining* (pp. 327–328).
Stamper, J. C., Lomas, D., Ching, D., Ritter, S., Koedinger, K. R., & Steinhart, J. (2012). The rise of the super experiment. In *Proceedings of the 5th international conference on educational data mining* (pp. 196–199).
Su, J., Tseng, S., Lin, H., & Chen, C. (2011). A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments. *User Modeling and User-Adapted Interaction, 21*(1–2), 5–49.
Sudol, L. A., Rivers, K., & Harris, T. K. (2012). Calculating probabilistic distance to solution in a complex problem solving domain. In *Proceedings of the 5th international conference on educational data mining* (pp. 144–147).
Sudol-DeLyser, L. A., & Steinhart, J. (2011). Factors impacting novice code comprehension in a tutor for introductory computer science. In *Proceedings of the 4th international conference on educational data mining* (pp. 365–366).

Sun, X. (2012). Finding dependent test items: an information theory based approach. In *Proceedings of the 5th international conference on educational data mining* (pp. 222–223).

Surpatean, A., Smirnov, E., & Manie, N. (2012). Similarity functions for collaborative master recommendations. In *Proceedings of the 5th international conference on educational data mining* (pp. 230–231).

Sweet, S. J., & Rupp, A. A. (2012). Using the ECD framework to support evidentiary reasoning in the context of a simulation study for detecting learner differences in epistemic games. *Journal of Educational Data Mining, 4*(1), 183–223.

Tabandeh, Y., & Sami, A. (2010). Classification of tutor system logs with high categorical features. In *Proceedings of the KDD 2010 cup 2010 workshop: Knowledge discovery in educational data* (pp. 54–61).

Tan, L. (2012). Fit-to-model statistics for evaluating quality of Bayesian student ability estimation. In *Proceedings of the 5th international conference on educational data mining* (pp. 224–225).

Thai-Nghe, N., Drumond, L., Horvath, T., & Schmidt-Thieme, L. (2011). Multi-relational matrix factorization models for predicting student performance. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 1–10).

Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization models for forecasting student performance. In *Proceedings of the 4th international conference on educational data mining* (pp. 11–20).

Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. In *Proceedings of the KDD 2010 cup 2010 workshop: Knowledge discovery in educational data* (pp. 13–23).

Trivedi, S., Pardos, Z. A., Sárközy, G. N., & Heffernan, N. T. (2011). Spectral clustering in educational data mining. In *Proceedings of the 4th international conference on educational data mining* (pp. 129–138).

Trivedi, S., Pardos, Z. A., Sárközy, G. N., & Heffernan, N. T. (2012). Co-clustering by bipartite spectral graph partitioning for out-of-tutor prediction. In *Proceedings of the 5th international conference on educational data mining* (pp. 33–40).

Tsuruta, S., Knauf, R., Dohi, S., Kawabe, T., & Sakurai, Y. (2012). An intelligent system for modeling and supporting academic educational processes. In A. Peña-Ayala (Ed.), *Intelligent and adaptive educational-learning systems: achievements and trends, smart innovation, systems and technologies* (pp. 469–496). Heidelberg: Springer.

Vialardi-Sacin, C., Bravo-Agapito, J., Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In *Proceedings of the 2nd international conference on educational data mining* (pp. 190–199).

Vialardi, C., Chue, J., Barrientos, A., Victoria, D., Estrella, J., Peche, J. P., & Ortigosa, A. (2010). A case study: data mining applied to student enrollment. In *Proceedings of the 3rd international conference on educational data mining* (pp. 333–334).

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., et al. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction, 21*(1–2), 217–248.

Vlahos, G. E., Ferratt, T. W., & Knoepfle, G. (2004). The use of computer-based information systems by German managers to support decision making. *Journal of Information & Management, 41*(6), 763–779.

Von-Davier, A. A. (2011). Quality control and data mining techniques applied to monitoring scaled scores. In *Proceedings of the 4th international conference on educational data mining* (pp. 353–354).

Vuong, A., Nixon, T., & Towle, B. (2011). A method for finding prerequisites within a curriculum. In *Proceedings of the 4th international conference on educational data mining* (pp. 211–215).

Wang, Q. Y., Kehrer, P, Pardos, Z. A., & Heffernan, N. (2011). Response tabling - a simple and practical complement to knowledge tracing. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 31–40).

Wang, Y., & Beck, J. E. (2012). Using student modeling to estimate student knowledge retention. In *Proceedings of the 5th international conference on educational data mining* (pp. 200–203).

Wang, Y., & Heffernan, N. T. (2011). Towards modeling forgetting and relearning in ITS: preliminary analysis of ARRS data. In *Proceedings of the 4th international conference on educational data mining* (351–352).

Wang, Y., & Heffernan, N. T. (2012). Leveraging first response time into the knowledge tracing model. In *Proceedings of the 5th international conference on educational data mining* (pp. 176–179).

Wang, Y., Heffernan, N. T., & Beck, J. E. (2010). Representing student performance with partial credit. In *Proceedings of the 3rd international conference on educational data mining* (pp. 335–336).

Wang, Y. H., & Liao, H. C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications, 38*(6), 6480–6485.

Warnakulasooriya, R., & Galen, W. (2012). Categorizing students' response patterns using the concept of fractal dimension. In *Proceedings of the 5th international conference on educational data mining* (pp. 214–215).

Wauters, K., Desmet, P., & van den Noortgate, W. (2011a). Acquiring item difficulty estimates: a collaborative effort of data and judgment. In *Proceedings of the 4th international conference on educational data mining* (pp. 121–127).

Wauters, K., Desmet, P., & van den Noortgate, W. (2011b). Monitoring learners proficiency: weight adaptation in the elo rating system. In *Proceedings of the 4th international conference on educational data mining* (pp. 247–252).

Wijaya, T. K., & Prasetyo, P. K. (2010). Knowledge tracing with stochastic method. In *Proceedings of the KDD 2010 cup 2010 workshop: Knowledge discovery in educational data* (pp. 51–53).

Witten, I. H., & Frank, E. (2000). *Data mining: practical machine learning tools and techniques with Java Implementations*. San Francisco: Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). California: Morgan Kaufmann.

Worsley, M, & Blikstein, P. (2011). What's an expert? using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In *Proceedings of the 4th international conference on educational data mining* (pp. 235–239).

Wu, X., Kumar, V., Quinlan, J. S., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Journal of Knowledge and Information Systems, 14*(1), 1–37.

Wu, Z., & Li, C. H. (2007). L0-constrained regression for data mining. In W. K. Ng, M. Kitsuregawa, J. Li, & K. Chang (Eds.), *Advances in knowledge discovery and data mining, Lecture Notes in Computer Science* (pp. 981–988). Heidelberg: Springer. 4426.

Xiong, W., Litman, D., & Schunn, C. (2010). Assessing reviewers' performance based on mining problem localization in peer-review data. In *Proceedings of the 3rd international conference on educational data mining* (pp. 211–220).

Xiong, X., Pardos, Z. A., & Heffernan, N. (2011). An analysis of response time data for improving student performance prediction. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 103–108).

Xu, B., & Recker, M. (2010). Peer production of online learning resources: a social network analysis. In *Proceedings of the 3rd international conference on educational data mining* (pp. 315–316).

Xu, B., & Recker, M. (2011). Understanding teacher users of a digital library service: a clustering approach. *Journal of Educational Data Mining, 3*(1), 1–28.

Xu, Y., & Mostow, J. (2011a). Logistic regression in a dynamic bayes net models multiple subskills better! In *Proceedings of the 4th international conference on educational data mining* (pp. 337–338).

Xu, Y., & Mostow, J. (2011b). Using logistic regression to trace multiple subskills in a dynamic bayes net. In *Proceedings of the 4th international conference on educational data mining* (pp. 241–245).

Xu, Y., & Mostow, J. (2012). Comparison of methods to trace multiple subskills: Is LR-DBN best? In *Proceedings of the 5th international conference on educational data mining* (pp. 41–48).

Yanto, I. T. R., Herawan, P., Herawan, T., & Deris, M. M. (2012). Applying variable precision rough set model for clustering student suffering study's anxiety. *Expert Systems with Applications, 39*(1), 452–459.

Yoo, J. S., & Cho, M. H. (2012). Mining concept maps to understand university students' learning. In *Proceedings of the 5th international conference on educational data mining* (pp. 184–187).

Yu, H. F., Lo, H. Y., Hsieh, H. P., Lou, J. K., G.McKenzie, T., Chou, J. W., Chung, P. H., Ho, C. H., Chang, C. F., Wei, Y. H., Weng, J. Y., Yan, E. S., Chang, C. W., Kuo, T. T., Lo, Y. C., Chang, P. T., Po, C., Wang, C. Y., Huang, Y. H., Hung, C. W., Ruan, Y. X., Lin, Y. S., Lin, S. Lin, H. T., & Lin, C. J. (2010). Feature engineering and classifier ensemble for KDD cup 2010. In *Proceedings of the KDD 2010 cup 2010 workshop knowledge discovery in educational data* (pp. 1–12).

Yudelson, M. V., & Brunskill, E. (2012). Policy building – an extension to user modeling. In *Proceedings of the 5th international conference on educational data mining* (pp. 188–191).

Yudelson, M. V., Pavlik, Jr., P. I., & Koedinger, K. R. (2011). Towards better understanding of transfer in cognitive models of practice. In *Proceedings of the 4th international conference on educational data mining* (pp. 373–374).

Yudelson, M., Brusilovsky, P., Mitrovic, A., & Mathews, M. (2010). Using numeric optimization to refine semantic user model integration of educational systems. In *Proceedings of the 3rd international conference on educational data mining* (pp. 221–230).

Zafra, A., Romero, C., & Ventura, S. (2011). Multiple instance learning for classifying students in learning management systems. *Expert Systems with Applications, 38*(12), 15020–15031.

Zapata-Gonzalez, A., Menendez, V. H., Prieto-Mendez, M. E. & Romero, C. (2011). Using data mining in a recommender system to search for learning objects in repositories. In *Proceedings of the 4th international conference on educational data mining* (pp. 321–322).

Zimmermann, J., Brodersen, K. H., Pellet, J. P., August, E., & Buhmann, J. M. (2011). Predicting graduate-level performance from undergraduate achievements. In *Proceedings of the 4th international conference on educational data mining* (pp. 357–358).

Zorrilla, M., García-Saiz, D., & Balcázar, J. L. (2011). Towards parameter-free data mining: mining educational data with yacaree. In *Proceedings of the 4th international conference on educational data mining* (pp. 363–364).