Applying Data Mining in Moodle

Cristóbal Romero Morales

(cromero@uco.es)

Department of Computer Sciences and Numerical Analysis. University of Córdoba



Outline

 Introduction to EDM
 Exporting Moodle data
 Preprocessing Moodle data
 Association Rule Mining in Moodle
 Classification and Clustering in Moodle

Introduction to EDM

Introduction

- The development of web-based educational systems has been rising exponentially in the recent years.
 - These systems produce information of high educational value, but usually so abundant that it is impossible to analyze it manually.
 - Tools to automatically analyze this kind of data are needed.
- Educational institutions have information systems that store plenty of interesting information.
 - This available information can be used to improve Strategic Planning of these institutions. In this case, tools to analyze that data automatically are also needed.



What do we call it?

- Statistics
- Machine Learning
- Data mining
- Knowledge Discovery in Data
- Business Analytics/Intelligent
- Data Analytics
- Big Data

• ?

Data Science

Same Core Idea: Finding Useful Patterns in Data

Different Emphasis

"In god we trust, all others must bring Data" William Edwards Deming (1900-1993)

Introduction What is EDS?

Educational Data Science (EDS) that only works with data gathered from educational environments/setting for solving educational problems.

Educational Data Science (EDS) is a multidisciplinary domain (computer science, education, statistics) with several related communities:



Introduction What is EDM?

Educational data mining (EDM) is the application of data mining techniques to educational environments.



Introduction Multidisciplinary domain

Educational data mining (EDM) is a multidisciplinary domain that is an intersection of 3 domains: computer science, education, statistics.



Introduction Other areas closely related to EDM

Learning analytics

□ The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.

Academic analytics

□Bussiness intelligence applied to institutional academic data.

Introduction Background on EDS

1991-3	First Papers about prediction
2000	Specific Workshops
2005	Data Mining in e-learning Book
2008	First EDM conference
2010	Handbook on EDM
2011	First LAK conference
2014	First L@S conference

Process and actors

The Lifecycle of Educational Data Science:



Educational Data Types of Educational Environments



Educational Data Characteristics

- The information come from different sources of data.
- There are a lot of incomplete and loss data because not all students carry out all the activities.
- User/Students are clearly identified.
- There is a great number of available instances and attributes that may required tasks of filtering for selecting the most important.
- Educational data have different level of granularity.
- Some transformation such as discretization of number are normally used for improving the comprehensibility of data and the obtained models.

Educational Data DM Tecnique used for each type of data

• Different types of data and DM techniques used:

Type of dataDM TechniqueRelational dataRelational data miningTransactional dataClassification, clustering, association rule mining, etc.Temporal, sequence and time series dataSequential data miningText dataText miningMultimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining		
Relational dataRelational data miningTransactional dataClassification, clustering, association rule mining, etc.Temporal, sequence and time series dataSequential data miningText dataText miningMultimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining	Type of data	DM Technique
Transactional dataClassification, clustering, association rule mining, etc.Temporal, sequence and time series dataSequential data miningText dataText miningMultimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining	Relational data	Relational data mining
Temporal, sequence and time series dataSequential data miningText dataText miningMultimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining	Transactional data	Classification, clustering, association rule mining, etc.
Text dataText miningMultimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining	Temporal, sequence and time series data	Sequential data mining
Multimedia dataMultimedia data miningWorld Wide Web dataWeb content/structure/usage mining	Text data	Text mining
World Wide Web dataWeb content/structure/usage mining	Multimedia data	Multimedia data mining
	World Wide Web data	Web content/structure/usage mining

EDS Publications Books

- <u>Data Mining in E-Learning</u>.
 C. Romero & S. Ventura (Eds).
 Editorial WIT Press, 2006.
- Handbook of Educational Data Mining.
 C.Romero, S. Ventura,
 M. Pechenizky, R. Baker. (Eds).
 Editorial CRC Press, Taylor & Francis Group. 2010.
- <u>Education Data Mining: Applications and Trends</u>.
 A. Peña-Ayala (Eds).
 Springer, SCI Vol. 524, 2014





Educational Data Mining

andro Peña-Ayala Editor

EDS Publications Books

 <u>Learning Analytics: From research to practice</u> J.A. Larusson, B. White (Eds).
 Springer, SCI Vol. 524, 2014



2 Springer

 Data Mining and Learning Analytics: Applications in Educational Research.
 S. ElAtia, D. Ipperciel, O.R. Zaïane.
 Wiley, 2016



EDS Publications Surveys/Reviews

- C. Romero & S. Ventura. Educational Data Mining: A survey from 1995 to 2005. Expert Systems with Applications 33:1, pp. 135-146, 2007.
- C. Romero, S. Ventura. 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, pp. 601 – 618. 2010.
- Karen Cator. Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics. Report of the U.S. Office of Educational Technology. 2012.
- C. Romero, S. Ventura. Data Mining in Education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Volume 3, Issue 1, pages 12–27, 2013.
- C. Romero, S. Ventura. Educational Data Science In MOOC. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Volume 7, Issue 1, pages 1–12, 2017.

DM Software



Weka is one of the most popular software packages for Data Mining

http://www.cs.waikato.ac.nz/~ml/weka/



This is a very popular DM tool, developed in Java

http://rapidminer.com



R is a programming language that was initially created to perform statistics, but it has also used in DM

https://www.r-project.org/

Specific EDS Software

ΤοοΙ	Objective	Reference
WUM tool	To extract patterns useful for evaluating on-line courses.	(Zaïane and Luo, 2001)
EPRules	To discover prediction rules to provide feedback for courseware authors.	(Romero et al., 2004)
GISMO/CourseVis	To visualize what is happening in distance learning classes.	(Mazza and Milani, 2004)
TADA-ED	To help teachers to discover relevant patterns in students' online exercises.	(Merceron and Yacef, 2005)
O3R	To retrieve and interpret sequential navigation patterns.	(Becker et al., 2005)
Synergo/ColAT	To analyze and produce interpretative views of learning activities.	(Avouris et al., 2005)
LISTEN Mining tool	To explore huge student-tutor interaction logs.	(Mostow et al., 2005)
MINEL	To analyze the navigational behavior and the performance of the learner.	(Bellaachia and Vommina, 2006)
LOCO-Analyst	To provide teachers with feedback on the learning process.	(Jovanovic et al., 2007)
Measuring tool	To measure the motivation of online learners.	(Hershkovitz and Nachmias, 2008)
DataShop	To store and analyze click-stream data, fine-grained longitudinal data generated	(Koedinger et al., 2008)
	by educational systems.	
Decisional tool	To discover factors contributing to students' success and failure rates.	(Selmoune and Alimazighi, 2008)
CIECoF	To make recommendations to courseware authors about how to improve courses.	(Garcia et al., 2009)
SAMOS	Student activity monitoring using overview spreadsheets.	(Juan et al., 2009)
PDinamet	To support teachers in collaborative student modeling.	(Gaudioso et al., 2009)
AHA! Mining Tool	To recommend the best links for a student to visit next.	(Romero et al., 2009)
EDM Visualization Tool	To visualize the process in which student solve procedural problem in logic.	(Johnson and Barnes, 2010)
Meerkat-ED	To analyze participation of students in discussion forums using social network	(Rabbany et al. 2011)
	analysis techniques.	
E-learning Web Miner	To discover student's behavior profiles and models about how they work in virtual	(García-Saiz and Zorrilla, 2011)
	courses.	
MMT tool	To facilitate the execution of all the steps in the data mining process of Moodle	(Pedraza-Perez et al., 2011)
	data for newcomers.	

Specific Moodle EDS Software

	Tool	Free	Integr.	Language	Vis.	Preproc.	Supervi.	Unsupervi.
	CoSyLMSAnalytics	No	No	VisualBasic	No	No	No	Yes
	ViMoodle	No	No	Java	Yes	No	No	No
	CIECoF	No	No	Java	No	No	No	Yes
	Meerkat-ED	No	No	Java	Yes	No	No	Yes
	ММТ	No	No	Java	No	Yes	Yes	Yes
	DRAL	No	No	Java	No	No	Yes	No
	GISMO	Yes	Yes	PHP	Yes	No	No	No
	http://gismo.sourceforge.net/							
	SNAPP	Yes	Yes	JavaScript	Yes	No	No	No
	AAT	No	No	PHP	No	No	No	No
	MOClog	Yes	Yes	PHP	Yes	No	No	No
	http://moclog.ch/							
	E-learningWebMiner	No	No	Java	Yes	No	No	Yes
	CVLA	No	Yes	Phyton	Yes	No	Yes	No
	IntelliBoard.net	No	Yes	PHP	Yes	No	No	No
	http://intelliboard.net/							
	SmartKlass	Yes	Yes	PHP	Yes	No	No	No
	http://klassdata.com/							
	Engagement Analytics	Yes	Yes	PHP	Yes	No	No	No
rg/	plugins/browse.php?list=set&id	d=20						
	Analytics Graphs	Yes	Yes	PHP	Yes	No	No	No
org	/plugins/block_analytics_graph	IS						

https://moodle.c

https://moodle.

Exporting Moodle data



Back up course content

■Administracion -> Copia de seguridad

preferencias personales

	novación educativa y Formación del P.D.I. las TIC Programación Docente	campus virtual enseñanza virtual y laboratorios tecnológicos	Contacta Idioma Salir
/ ► Innovación educativa y Formación del P.	D.I. 🕨 Mis asignaturas en este Centro 🕨 Plan de Formación del P	DI (2015-2016) 🕨 Learning Analytics: Aplicación de Técnicas de Mine 🕨 Copia de seguridad	
lavegación	Configuración de la conia de seguridad	1. Ajustes iniciales ► 2. Ajustes del esquema ► 3. Confirmación y revisión ► 4. Realizar copia de seguridad ► 5. Complet	ar
 novación educativa y Formación del P.D Mi área personal Panel de mensajes personales y notificaciones Mi información personal Asignatura actual Learning Analytics: Aplicación 	I. IMS Common Cartridge 1.1 Incluir participantes inscritos Hacer anónima la información de participante Incluir asignaciones de rol de participante Incluir actividades y contenidos	 × 3a × 3a × 3a ∞ 	
Técnicas de Mine) Participantes Mis asignaturas en este Centro Asignaturas dministración Administración de la asignatura Attivar edición Modificar ajustes) Gestión de participantes Tiltros	Incluir bloques Incluir filtros Incluir comentarios Incluir comentarios Incluir insignias Incluir eventos del calendario Incluir detalles del grado de avance del participante Incluir detalles del grado de avance del participante Incluir borso de la asignatura Incluir historial de calificaciones Incluir banco de preguntas		
 Informes Calificationes Copiar desde otra asignatura Copia de seguridad Restaurar Restablecer Banco de preguntas Cambiar mi rol a Mis ajustes de información y 		Cancelar	Saltar al último paso Siguiente

Export gradebook

Administración -> Calificaciones Administracion de calificaciones -> Exportar

UNVERTIDAD UNVERTIDAD DE WALLARD	ración educativa y Formación del P.D.I. IC Programación Docente		campus virtual enserianza	virtual y tecnológicos		ovación educativa y Formación del P.D.I. TTC Programación Docente	campus virtual enseitance virtual y laboratorios tecnológicos	Contacta Idioma Salir
CV ▶ Innovación educativa y Formación del P.D.I.	▶ Mis asignaturas en este Centro ▶ Plan de Forma	ción del PDI (2015-2016) ► I	Learning Analytics: Aplicación de Técnicas de Mine	► Administración de calificaciones ► Infor	e de CV ⊫ Innovación educativa y Form	mación del P.D.I. ▶ Mis asignaturas en este Centro ▶ Plan da	e Formación del PDI (2015-2016) 🕨 Learning Analytics: Aplicación de Técnicas de M	line 🕨 Administración de calificaciones 🕨 Exportar 🕨
					Hoja de cálculo Excel			
Navegación 🔍	Vista Ajuste Escalas no numério	cas Competencias	Letras Importar Exportar		Navegación	Vista Aiuste Escalas no numéricas	Competencias Letras Importar Exportar	
Innovación educativa y Formación del P.D.I.	Informe del evaluador Historial	de calificación Inform	e de competencias Vista simple Info	me al estudiante	Innovación educativa v	Alfil Hoja de cálculo OpenOffice Arch	ivo en texto simple sin formato Hoja de cálculo Excel Archivo 3	XML
 Panel de mensajes personales y notificaciones 	Informe del evaluador				Formación del P.D.1.	Exportar a Hoja de cálculo Ex	cel	
Mi información personal					Panel de mensajes persona	nales		
 Asignatura actual 	Número de participantes:21	/21			y induitaciónes	☐ ▼Ítems de calificación a incluir —		Contraer t
Learning Analytics: Aplicación de Técnicas de Mine Participantes	Nombre: Todos A B C D E F G	HIJKLMNÑ	İ O P Q R S T U V W X Y Z		 Asignatura actual 	Entrega		
Mis asignaturas en este Centro	Apellido/s: Todos A B C D E F	GHIJKLMN	ÑOPQRSTUVWXYZ		 Learning Analytics: Aplicación de Técnicas 	as Prueba Test	™ 17	
Asignaturas		tooning tools			de Mine	Total de la asignatura		
Administración	Apellido/s 📥 Nombre	Learning Analyt	ics: 🖻	🕴 Total de la asignatura 🚔 🥖	Participantes		Seleccionar todos/ninguno	
Administración de calificaciones Informe del evaluador Historial de calificación	Alang Tak Bark	Satisfactori	0 95,00	94,12	 Mis asignaturas en esce Centro Asignaturas 	Opciones de los formatos de export Incluir retroalimentación en la exportación		
 Informe de competencias Vista simple Informe al estudiante 	Balmonia Marilea Maria Mittera	2 Satisfactori	0 100,00	99,02	Administración	Excluir participantes suspendidos (?) Forma de mostrar exportación de	 ☑ ☑ Real □ Porcentaje □ Letra	
Ajuste Importar Evontar	Biogue Para Oda	2 Satisfactori	o 0,00	0,98	Administración de calificaciones GI Informe del evaluador	Puntos decimales en la exportación de calificaciones	2 🗸	
 ▲ Alfil ▲ Hoja de cálculo OpenOffice ▲ Hoja de asistencia (PDF) 	Dollar. Androis Canador	Satisfactori	0 80,00	79,41	Historial de calificación Informe de competenci Vista simple	cias	_	
 Archivo en texto simple sin formato Hoja de cálculo Excel 	Califord Royal	2			 Informe al estudiante Ajuste 		Descargar	
📩 Archivo XML	A Index New Workley	2			 Importar Exportar Alfi 			
Competencias Il Escalas no numéricas Administración de la asignatura	Mandon Modele	Satisfactori	0 80,00	79,41	☆ Hoja de cálculo OpenOffice			
 Cambiar mi rol a Mis ajustes de información y preferencias personales 	Maronan Galileon Dece Senti	2			☆ Hoja de asistencia (PDF)			
	Bronneite gen	eral Satisfactori	D 79,17	76,40	simple sin formato			
		Caticfactori	100.00	00.02				

Export reports

Administracion de asignatura -> Informes

Moodle allows instructors to request reports telling which resources and activities of a course have been accessed, when, and by whom. Moodle produces several kinds of reports:

- Logs generates a filtered report showing information about a particular activity or student.
- Activity report generates a simple unfiltered report showing all activity in the course that you can sort by column header.
- Course participation provides a sortable list showing all class members, with details about a particular resource or activity. You can see who has viewed a resource or submitted an activity. From this screen, instructors can also send a message to all students, or only to those students who have not completed an activity.

Export reports (logs)

■Administracion de asignatura -> Informes -> Registros

	(NNOVAC Aulas TIC	ión educati : Programaci	va y Formación del P. ión Docente	.D.I.	campus virtual		Contacta	a Idioma Salir		
CV 🕨 Innovación educativa y Formación	del P.D.I.	► Mis asignat	turas en este Centro 🕨 Pla	an de Formación del PDI	(2015-2016) 🕨 Learning Analytics: Aplicación de	Técnicas de Mine 🕨	Informes 🕨 Registros			
Navegación	avegación 📮 Learning Analytics: Aplicación de Técnicas de Minería y Análisis de Datos en Educación. (2015-16) 🗸 Número de participantes 🗸 Todos los días 🗸									
Innovación educativa y Formación del P.D.I. Mi área personal Página: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 (Siguier						 Registros desd 16 17 18 (Siguiente) 	e 26-09-2015 Conseguir estos regis Mostrar todos	stros		
 Mi información personal Asignatura actual 		Hora	Nombre completo del participante	Participante afectado	Contexto del evento	Componente	Nombre evento	Descripción	Origen	Dirección IP
 Learning Analytics: Aplicac de Técnicas de Mine Participantes Mis asignaturas en este Centro 	ión	14 de dic, 10:07	Remers Realist Cratilist		Asignatura: Learning Analytics: Aplicación de Técnicas de Minería y Análisis de Datos en Educación. (2015-16)	Sistema	Asignatura vista	The user with id '6275' viewed the course with id '610'.	web	150.214.118.98
Asignaturas Administración		13 de dic, 20:27	Remen Media Grafilial	Barraen (Bendisc Cristilisa)	Asignatura: Learning Analytics: Aplicación de Técnicas de Minería y Análisis de Datos en Educación. (2015-16)	Informe al estudiante	Vista del informe de calificación del participante	The user with id '6275' viewed the user report in the gradebook.	web	92.185.123.199
Administración de la asignatura Activar edición Modificar ajustes		13 de dic, 19:46	Ramann Manaka Graebhai		Asignatura: Learning Analytics: Aplicación de Técnicas de Minería y Análisis de Datos en Educación. (2015-16)	Sistema	Asignatura vista	The user with id '6275' viewed the course with id '610'.	web	92.185.123.199
 Filtros Informes Registros 		13 de dic, 19:46	Remers Wenders Credition		Foro: Foro de dudas, sugerencias y colaboración	Foro	Discussion viewed	The user with id '6275' has viewed the discussion with id '3986' in the forum with course module id '14927'.	web	92.185.123.199
 Registros activos Informe de actividad Participación en la asignat 	ura	13 de dic, 19:46	Remerc Menders Credited		Foro: Foro de dudas, sugerencias y colaboración	Foro	Módulo de asignatura visto	The user with id '6275' viewed the 'forum' activity with course module id '14927'.	web	92.185.123.199
Calificaciones Competencias Copiar desde otra asignatura https://formacionpdi.cv.uma.es/course	e/view.ph	13 de dic, 19:46 np?id=610	Namen Medite Gratiliai		Asignatura: Learning Analytics: Aplicación de Técnicas de Minería y Análisis de Datos en Educación. (2015-16)	Sistema	Asignatura vista	The user with id '6275' viewed the course with id '610'.	web	92.185.123.199

Download Quiz Data

Pulsar sobre el enlace del Test. Administracion de la prueba de conocimiento -> Resultados -> Respuestas detalladas

	in educativa y Formación del P.D.I. Programación Docente	campus virtual enseñanza virtual y Laboratorios tecnológicos	Contacta Idioma Salir
CV 🕨 Innovación educativa y Formación Respuestas detalladas	del P.D.I. ► Mis asignaturas en este Centro ► Plan de F	rmación del PDI (2015-2016) ► Learning Analytics: Aplicación de Técnicas de Mine ► Tema inicial)	▶ Prueba Test ▶ Resultados ▶
Navegación 🔍	Prueba Test		
Innovación educativa y Formación del P.D.I. Mi área personal	Intentos: 1		▼ Contraer todo
 Panel de mensajes personales y notificaciones 	▼Qué incluir en el informe Los intentos de	todos los participantes que han hecho intentos de resolver la prueba	
 Mi información personal Asignatura actual 	Los intentos que hay	En curso Atrasado Finalizado Nunca presentó	
 Learning Analytics: Aplicación de Técnicas de Mine 	✓ Mostrar opciones	·	
Participantes Tema inicial	Tamaño de página	30	
 Mis asignaturas en este Centro 	Mostrar el/la	🗌 texto de la pregunta 🛛 respuesta 🗌 respuesta correcta	
Asignaturas		Mostrar informe	
Administración 🛛 🗖			
 Administración de la prueba de conocimiento Modificar ajustes 	Descargar d	Sólo se permite un intento por participante en esta prueba de conocimiento tos de tabla como Archivo de texto con valores separados por comas	Descargar
 Evitar participación de grupos 			
 Evitar participación de participante 			
Modificar la prueba de conocimiento	Romero Moral Revisión del inte	es Cristóbal Finalizado 0,00 nto	🗙 з
 Vista previa Resultados Calificaciones 		Seleccionar todos / Deseleccionar todos Borrar los intentos seleccionados	
 Respuestas detalladas 			
 Estadísticas Calificación manual 			
 Roles asignados localmente 			
Permisos			

Preprocessing Moodle Data

Preprocesing Data Introduction

Data Mining Process with Moodle data: ullet

Preprocess Data

Collect Moodle Usage Data



Apply Data Mining Algorithms

218

lass: mark (Non

Save ...

Type: Nominal Jnique: 0 (0%)

Apply

Visualize All

Log

Preprocesing Data Introduction

- The first step in any KDD process is the transformation of data into an appropriate form for the mining process.
- Data pre-processing in educational context is considered the most crucial phase in the whole educational data mining process, and it cantake more than half of the total time spent in solving the data mining problem.
- The data pre-processing phase typically consumes 60-80% of the time of the KDD process.

Preprocesing Data Introduction

 The main steps/tasks of the overall process of preprocessing educational data are:



• Example of gathering, data aggregation and integration:



Moodle's Data Base has more than 200 Tables:



32

• Using SQL to access Moodle's Tables:

SELECT COUNT(*) FROM mdl_quiz,mdl_quiz_grades WHERE mdl_quiz_grades.userid= " +userid+ " and mdl_quiz.course= " + id + " and mdl_quiz.id = mdl_quiz_grades.quiz

MySQL Administrator	- root@localhost:3306						
File Edit View Tools Wind	low Help						
Server Information Service Control Startup Variables	Schema Tables Schema Indices Views Store moodle All tables of the moodle schema	d procedures					
	Table Name 🔺	Engine	Rows	Data length	Index length	Update time	
	mdl_assignment	MyISAM	186	170,2 kB	7 kB	2006-02-20 16:27:36	~
Mealth	mdl_assignment_submissions	MyISAM	6288	426,3 kB	251 kB	2006-02-21 21:48:52	
Enver Logs	mdl_backup_config	MyISAM	0	0 B	1 kB	2006-02-09 08:18:46	
Beplication Status	md_backup_courses	MyISAM	0	0 B	1 kB	2006-02-09 08:18:46	
	md_backup_files	MyISAM	0	73,1 kB	43 kB	2006-02-21 11:11:38	
🥥 Backup	mdl_backup_ids	MyISAM	0	95,5 kB	31 kB	2006-02-21 11:11:38	
Sestore	mdl_backup_log	MyISAM	0	0 B	1 kB	2006-02-09 08:18:46	
Catalogs	md_block	MyISAM	20	548 B	2 kB	2006-02-09 08:18:46	
	mdl_block_instance	MyISAM	1852	57,9 kB	48 kB	2006-02-22 11:52:48	
	md_block_rss_client	MyISAM	0	0 B	1 kB	2006-02-09 08:18:46	
Schemata	md_book	MyISAM	14	1 kB	2 kB	2006-02-09 08:18:46	
	md_book_chapters	MyISAM	34	11,5 kB	2 kB	2006-02-09 08:18:46	
	md_cache_filters	MyISAM	392	31,2 kB	25 kB	2006-02-22 10:21:04	
	mdl_cache_text	MyISAM	192	1,4 MB	186 kB	2006-02-22 12:00:04	
Corden		MUCAN	0	CAAD	01.D	2000 02 22 00 40.02	
information_schema moodle mysql rules_repository	Num. of Tables: 145 Details >>	Row	s: 2.32 eate Table	Edit Table	n: 165,3 MI	B Index Len: 103,8	MB

List of important tables in Moodle database about student interaction:

Name	Description
mdl_user	Information about all the users.
mdl_user_students	Information about all students.
mdl_log	Logs every user's action.
mdl_assignement	Information about each assignment.
mdl_assignment_submissions	Information about assignments submitted.
mdl_forum	Information about all forums.
mdl_forum_posts	Stores all posts to the forums.
mdl_forum_discussions	Stores all forum discussions.
mdl_message	Stores all the current messages.
mdl_message_reads	Stores all the read messages.
mdl_quiz	Information about all quizzes.
mdl_quiz_attempts	Stores various attempts at a quiz.
mdl_quiz_grades	Stores the final quiz grade.

EDM Data Data Cleaning

 Example of data cleaning by plotting data clusters and discovering outliers or rare/anormal students:



EDM Data **Data Cleaning**

Missing data is a common issue in education (usually appear when students have not completed or done all the activities in the course) and some possible solutions are:

- Students who have missing values can be removed.
- Whenever possible, these specific students may be contacted and asked (by the instructor) to complete the course.
- To codify missing/unspecified values by mapping incomplete values using for example the labels "?" (missing) and "null" (unspecified).
- To use a global constant to fill in the missing value or to use a substitute value, like the attribute mean or the mode. 36
EDM Data User and Session Identification

- Although user and session identification is not specific to education, it is especially relevant due to the longitudinal nature of student usage data.
- Computer-based educational systems provide user authentication (identification by login and password). So it is not necessary to do the typical user and session identification.
- It is also necessary to preserve student data anonymity/privacy but enabling that different pieces of information are linked to the same person. A common solution for it consists in using a number randomly or incrementally generated, like a user ID.

EDM Data Data Filtering

• Example of **filtering** at different levels of granularity and their relationship to the amount of data:



EDM Data Attribute Selection

 Example of Summary Table with a set of attributes selected per student in Moodle courses:

Name	Description
id_student	Identification number of the student.
id_course	Identification number of the course.
num_sessions	Number of sessions.
num_assigment	Number of assignments done.
num_quiz	Number of quizzes taken.
a_scr_quiz	Average score on quizzes
num_posts	Number of messages sent to the forum.
num_read	Number of messages read on the forum.
t_time	Total time used on Moodle.
t_assignment	Total time used on assignments.
t_quiz	Total time used on quizzes.
t_forum	Total time used on forum.
f_scr_course	Final score of the student obtained in the course.

EDM Data Data Transformation

• Example of **transformation** is Discretization:

Manual discretization has the user himself directly specifying the cut-off points. Example (Marks/Scores depend on the country):
 FAIL: if value is < 5
 PASS: if value is ≥ 5 and < 7
 GOOD: if value is ≥ 7 and < 9
 EXCELLENT: if value is ≥ 9

EDM Data Data transformation

 Example of derived attributes, which enables to create new attributes starting from the current ones:

Name	Description	
UserId	A unique identifier per user.	
	Percentage of correctly answered tests calculated	
Performance	as the number of correct tests divided by the total	
	number of tests performed).	
TimeDeading	Time spent on pages (calculated as the total time	
TimeReading	spent on each page accessed) in a session.	
NoPages	The number of accessed pages.	
	The time spent performing tests (calculated as	
TimeTests	the total time spent on each test).	
Motivation	Engaged / Disengaged.	

EDM Data Data transformation

• Example of Moodle Summary ARFF file:

- O X Moodle-Summary.arff - Bloc de notas Archivo Edición Formato Ver Ayuda @relation student_summarization @attribute id_student numeric @attribute id course numeric @attribute num_sessions {HIGH, MEDIUM, LOW} @attribute num_assignment {HIGH, MEDIUM, LOW} @attribute num_quiz {HIGH, MEDIUM, LOW} @attribute a_scr_course {FAIL, PASS, GOOD, EXCELLENT} @attribute num_posts {HIGH, MEDIUM, LOW} @attribute num_read {HIGH, MEDIUM, LOW} @attribute t_time {HIGH, MEDIUM, LOW} @attribute t_assignment {HIGH, MEDIUM, LOW} @attribute t_quiz {HIGH, MEDIUM, LOW} @attribute t_forum {HIGH, MEDIUM, LOW} @attribute f_scr_course {FAIL, PASS, GOOD, EXCELLENT} @data 1,88,LOW,MEDIUM,HIGH,FAIL,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL 2,88,LOW,MEDIUM,HIGH,FAIL,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,FAIL 3,88,LOW,LOW,LOW,?,LOW,LOW,LOW,LOW,LOW,FAIL 4,88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,LOW,LOW,MEDIUM,LOW,GOOD 5,88,HIGH,HIGH,GOOD,LOW,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,EXCELLENT 6,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL 7.88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,PASS 8,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,MEDIUM,LOW,LOW,FAIL 9,88,LOW,HIGH,PASS,LOW,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,PASS 10,88,LOW,HIGH,FAIL,LOW,LOW,LOW,LOW,LOW,LOW,LOW,FAIL 11,88,MEDIUM,HIGH,PASS,LOW,LOW,LOW,MEDIUM,MEDIUM,LOW,PASS

Association Rule Mining in Moodle

Introduction Association Rule Mining task

- Association Rule Mining (ARM) is one of the most popular and well-known data mining methods for discovering interesting relationships between variables in data
- The extracted knowledge has the form of rules, that can be applied on a fraction of the examples stored in the database.
- Types: frequent AR, rare AR, general AR, interesting AR...

IF (asignments > 10) **THEN** (grade > 6) support 80% confidence:90%

antecedent

consequent

Rule quality measure

Related task: Sequential pattern mining, subgroup discovery, etc.

Introduction Rare Association Rules

- **Rare Association Rules** also known as non-frequent, unusual, exceptional or sporadic rules are those that only appear infrequently even though they are highly associated with very specific data
- Rare itemsets are those that only appear together in very few transactions or some very small percentage of transactions in the database.
- They have low support and high confidence in contrast to general association rules which are determined by high support and a high confidence level.



Introduction

- **ARM** has been applied extensively in e-learning to discover frequent student-behavior patterns.
- However, RARM has been hardly applied to educational data, despite the fact that infrequent associations can be of great interest since they are related to rare but crucial cases. These rules could help the instructor to discover a minority of students who may need specific support with their learning process.
- The greatest reason for applying RARM in the field of EDM is the imbalanced nature of data in education in which some classes have many more instances than others.
- Furthermore, in applications like education, the minor parts of an attribute can be more interesting than the major parts; for example, students who fail or drop out are usually less frequent than those students who fare well.⁴⁶

Experimentation and Results Data

- In order to test the performance and usefulness of applying ARM and RARM to e-learning data, we have used student data gathered from the Moodle system.
- These data are from 230 students in 5 Moodle courses on computer science at the University of Córdoba about all activities that students perform on-line (e.g., assignments, forums and quizzes).
- This student usage data has been preprocessed in order to be transformed into a suitable format to be used by our data mining algorithms.

Experimentation and Results Summary Table

• We have created a summary table which integrates the most important information about the on-line activities and the final marks obtained by students in the courses.

Name	Description	Values
course	Identification number of the course.	C218, C94, C110, C111, C46
n_assigment	Number of assignments done.	ZERO, LOW, MEDIUM, HIGH
n_quiz	Number of quizzes taken.	ZERO, LOW, MEDIUM, HIGH
n_quiz_a	Number of quizzes passed.	ZERO, LOW, MEDIUM, HIGH
n_quiz_s	Number of quizzes failed.	ZERO, LOW, MEDIUM, HIGH
n_posts	Number of messages sent to the forum.	ZERO, LOW, MEDIUM, HIGH
n_read	Number or messages read on the forum.	ZERO, LOW, MEDIUM, HIGH
total_time_assignment	Total time spent on assignments.	ZERO, LOW, MEDIUM, HIGH
total_time_quiz	Total time spent on quizzes.	ZERO, LOW, MEDIUM, HIGH
total_time_forum	Total time spent on forum.	ZERO, LOW, MEDIUM, HIGH
mark	Final mark obtained by the student in the course.	ABSENT, FAIL, PASS, EXCELLENT

Experimentation and Results Imbalanced Attributes

• Due to the way their values are distributed, the course and mark attributes are clearly imbalanced, i.e., they have one or more values with a very low percentage of appearance.



Experimentation and Results Class Association Rules

- We performed a comparison between ARM and different RARM algorithms to discover **Rare Class Association Rules**.
- A Class Association Rule is a special subset of association rules with the consequent of the rule limited to a target class label (only one predefined item in our case Mark attribute).

Item1 \cap *item2* \cap ... \cap *Itemn* \rightarrow *Class*

- In our specific context, these rules are very useful for educational purposes, since they show any existing relationships between the activities that students perform using Moodle and their final exam marks.
- To obtain Class Association Rules we have modified ARM and RARM algorithms in order to obtain only those rules that have a single attribute (in our case, the mark attribute) in their consequent.

Experimentation and Results Parameters

 We evaluated the four different Apriori proposals with the following configuration parameters:

- Apriori-Frequent, setting the minimum support threshold at a very low value (0.05).
- Apriori-Infrequent, Apriori-Inverse and Apriori-Rare setting the maximum support at 0.1.

We also assigned the value 0.7 as the confidence threshold for all the algorithms.

Experimentation and Results Summary of Results

Comparison Table of ARM and RARM proposals:

Algorithm	# Freq. Itemsets	# UnFreq. Itemsets	# Rules	Avg Support/ ± Std Deviation	Avg Confidence/ ± Std Deviation
Apriori-Frequent	11562		788	0.162±0.090	0.717±0.211
Apriori-Infrequent		1067	388	0.058±0.060	0.863±0.226
Apriori-Inverse		3491	46	$0.056{\pm}0.070$	0.883±0.120
Apriori-Rare		5750	44	$0.050{\pm}0.080$	$0.885{\pm}0.108$

Examples of discovered rules

- Next, we show some examples of rules that were obtained using A) the ARM (Apriori) and B) RARM (Apriori-Rare) algorithms.
- For each rule, we show the antecedent and the consequent constructed, as well as some evaluation rule measures such as the support, the confidence and two different versions of the conditional support.

Experimentation and Results Rule Evaluation Measures

• Due to the imbalanced nature of our data, we use different versions of the conditional support [Zhang et al. 2009]:

- Traditional support:
$$Sup(A \rightarrow C) = \frac{n(A \cap C)}{N}$$

– Conditional support with respect to the mark attribute:

$$SupM(A \rightarrow Mark) = \frac{n(A \cap Mark)}{n(Mark)}$$

Conditional support with respect to the course attribute:

 $SupC(A \cap Course \to Mark) = \frac{n(A \cap Course \cap Mark)}{n(Course)}$

Examples of discovered rules

Rules extracted using the Apriori-Frequent algorithm.

Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	total_time_forum=HIGH	mark=PASS	0.24	/0.47	0.82
2	n_posts=MEDIUM AND n_read=MEDIUM AND n_quiz_a=MEDIUM	mark=PASS	0.13	/0.25	0.71
3	course=C110 AND n_assignment=HIGH	mark=PASS	0.14	0.52/0.27	0.89
4	total_time_quiz=LOW	mark=FAIL	0.21	/0.55	0.78
5	n_assignment=LOW	mark=FAIL	0.23	/0.60	0.70
6	n_quiz_a=LOW AND course=C218	mark=FAIL	0.18	0.51/0.47	0.83

Examples of discovered rules

• Rules extracted using the Apriori-Rare algorithm.

Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	n_quiz=HIGH AND n_quiz_a=HIGH	mark=EXCELLENT	0.045	/0.69	0.86
2	total_time_assignment=HIGH	mark=EXCELLENT	0.045	/0.69	0.86
3	n_posts=HIGH AND course=C46	mark=EXCELLENT	0.045	1.00/0.69	1.00
4	total_time_assignment=ZERO AND total_time_forum=ZERO AND total_time_quiz=ZERO]	mark=ABSENT	0.050	/0.76	0.78
5	n_posts=ZERO AND n_read=ZERO	mark=ABSENT	0.050	/0.76	0.78
6	n_quiz=ZERO AND course=C111	mark=ABSENT	0.050	0.88/0.76	1.00

Classification and Clustering in Moodle

Tasks Classification

- Identifying to which set of categories a new observation belongs on the basis of a training set containing observations (instances) whose category membership is known (supervised learning method).
- Example: Build a model to predict if a given student will pass or not from certain information.
 - To do that...
 - I have information about students previously graded as "pass" or "fail". Those examples can contain different kind of information.
 - I build a model using a classification algorithm.
 - The model allow us to predict if a new student, whose information is provided to the model, will pass or fail.

Tasks Clustering

- Grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups or clusters.
- Example: Defining group of similar students from the usage informacion taken of a virtual learning system
 - To do that...
 - We have a set of **unlabelled** data.
 - The cluster algorithm search similarities between data and defines group of students with similar features.
 - The final model includes the description of the resulting groups.

Introduction

- Mining data generated by students communicating using forum-like tools can help reveal aspects of their communication.
- The more students participate in the forum for a certain course, the more involved they will be in the subject matter of that course.
- Following this line, in this study we try to test whether or not there is a correlation between the participation of students in Moodle forums and their final course marks.

Background

- The use of data mining is a potential strategy for discovering and building alternative representations for the data underlying discussion forums.
- There is less published work on the use of data mining to predict student performance based on forum usage data.
- Furthermore, the use of clustering for classification has not yet been applied in an educational context.

Proposed Approach

- We propose to use a meta-classifier that uses a cluster for classification approach based on the assumption that each cluster corresponds to a class.
- For all cluster algorithms, the number of clusters generated is the same as the number of class labels in the dataset. We use this approach to test if student participation in forums is related to whether they pass or fail the course.

Proposed Approach Proposed classification via clustering approach



Description of the data used

- The dataset used in this work was gathered from a Moodle forum used by university students during a first-year course in computer engineering in 2011.
- We developed a new module for Moodle specifically to obtain a summary dataset file.

Student	nMessages	nThreads	nReplies	nWords	nSentences	nR
Reyes	3	0	3	67	3	
Gamaz	6	1	5	513	1	
Njama Sanano	1	1	0	17	2	
lovar Mittlina	2	0	2	43	2	
-						-

Description of the data used

Some forum statistics are:

Number of students	Number of messages	Number of threads	Number of replies
114	1014	81	933

• The variables relating to forum usage are:

Attribute	Description
nMessages	Number of messages sent per student
nThreads	Number of threads created per student
nReplies	Number of replies sent per student
nWords	Number of words written by student
nSentences	Number of sentences written by student
nReads	Number of messages read on the forum
tTime	Total time, in hours, spent on forum
aEvaluation	Average score of the messages
dCentrality	Degree centrality of the student
dPrestige	Degree prestige of the student
fMark	Final mark obtained by the student

- In the first experiment, we executed the following clustering algorithms provided by Weka for classification via clustering using all attributes: EM, FarthestFirst, Xmeans, sIB HierarchicalClusterer and SimpleKMeans.
- In the second experiment, we repeated all the previous executions using fewer attributes, based on the assumption that not all the available attributes are discriminative factors in the final marks.

- We apply a range of feature-selection algorithms. To rank the attributes, we counted the number of times each attribute was selected by each attribute-selection algorithm.
- We selected as the best attributes the first six attributes in the ranking, because these were selected by at least half of the algorithms.

Attribute	Frequency
dCentrality	9
nMessages	8
nReplies, nWords	7
dPrestige	6
aEvaluation	5
nSentences, nReads, nThreads	3
tTime	1

 The table shows the overall accuracy (rate of correctly classified students) using all the available attributes (A) and using only the six selected attributes (B).

Clustering algorithm	(A)	(B)
EM	0.842	0.894
FarthestFirst	0.526	0.535
HierarchicalClusterer	0.578	0.570
sIB	0.710	0.578
SimpleKMeans	0.666	0.640
Xmeans	0.666	0.640

 In the third experiment, we compared the accuracy of the previous classification via clustering approach with that of traditional classification algorithms by executing a representative number of classifications of different types: Rules-based algorithms, Trees-based algorithms, Functions-based algorithms and Bayes-based algorithms.

Algorithms	(A)	(B)
DTNB	0.859	0.833
JRip	0.833	0.815
NNge	0.842	0.807
Ridor	0.833	0.842
ADTree	0.859	0.842
J48	0.824	0.807
LADTree	0.868	0.850
RandomForest	0.850	0.833
Logistic	0.859	0.850
MultilayerPerceptron	0.842	0.868
RBFNetwork	0.868	0.886
SMO	0.868	0.886
BayesNet	0.877	0.842
NaiveBayesSimple	0.859	0.894

• Finally, we show the cluster centroids for the EM algorithm when using the six selected attributes that have yielded the best accuracy.

Attributes	Cluster 0	Cluster 1
nMessages	1.2199	14.8905
nReplies	1.1599	13.6718
nWords	18.4599	668.8039
aEvaluation	0	0.7751
dCentrality	0.0011	0.1565
dPrestige	0	0.1021