

Educational Data Mining and DataShop

Ryan S.J.d. Baker
Kenneth R. Koedinger

EDM definition

- “the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in.” (Baker, in press)

EDM methods

- Often differ from regular DM methods
 - Attempt to exploit multiple levels of meaningful hierarchy
 - Keystroke, answer, session, student, classroom, school
 - Need to account for specific types of non-independence in data

(Some) Classes of method

- Prediction
- Clustering
- Relationship Mining
- Discovery with Models
- Distillation of Data For Human Judgment

Particularly prominent in EDM (as compared to DM in general)

- Prediction
- Clustering
- Relationship Mining
- **Discovery with Models**
- **Distillation of Data For Human Judgment**
- As in other areas of computational science

Prediction

- Develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables)
- Classification, regression, knowledge tracing
- Very prominent in EDM
 - EDM2009 best paper
 - Ritter et al
 - EDM2009 best student paper
 - Hershkovitz et al

Clustering

- Find points that naturally group together, splitting full data set into set of clusters
- Related: finding the factors that split the space of data features
 - Example: Principle Component Analysis
- Moderately prominent in EDM

Relationship Mining

- Discover relationships between variables in a data set with many variables
 - Association rule mining
 - Correlation mining
 - Sequential pattern mining
 - Causal data mining
- Prominent in EDM

Discovery with Models

- Pre-existing model (developed with EDM prediction methods... or clustering... or knowledge engineering...)
- Applied to data and used as a component in another analysis
- Moderately prominent in EDM
 - AIED2009 best paper nominee
 - Baker et al

Distillation of Data for Human Judgment

- Making complex data understandable by humans to leverage their judgment
- Moderately prominent in EDM

EDM track schedule (changed)

- Tuesday 10am
 - Bayesian Knowledge Tracing and Discovery with Models
- Tuesday 11am
 - Item Response Theory and Learning Factor Analysis
 - Ken Koedinger
- Tuesday 2:15pm
 - Prediction: Classifiers and Regressors
- Wednesday 11am
 - Principal Component Analysis
 - Geoff Gordon

Where does the data come from?

Educational Software

- Computer tutors and other educational software
 - Fine grained, longitudinal, often across contexts
- Instrumented to give logs of every “transaction” between student and software
 - Transaction = entering an answer, requesting help – some semantic “action”
- Sometimes also instrumented to give logs of mouse movements and keyboard actions (e.g. de Vicente & Pain, 2008)
- Sometimes even physical sensors (e.g. Arroyo et al, 2009)

Other data sources

- Other data sources
 - Records of online courses (e.g. WebCT, Moodle) (e.g. Romero et al, 2008; Riley et al, 2009)
 - District or university-level student records
 - Example: www.icpsr.umich.edu/IAED
 - Group collaboration data (e.g. Kay et al, 2008)
 - Records of library borrowing
 - Nice paper on this at EDM2009

Often annotated...

- Quantitative Field Observations



Often annotated...

- Text Replay Labelings

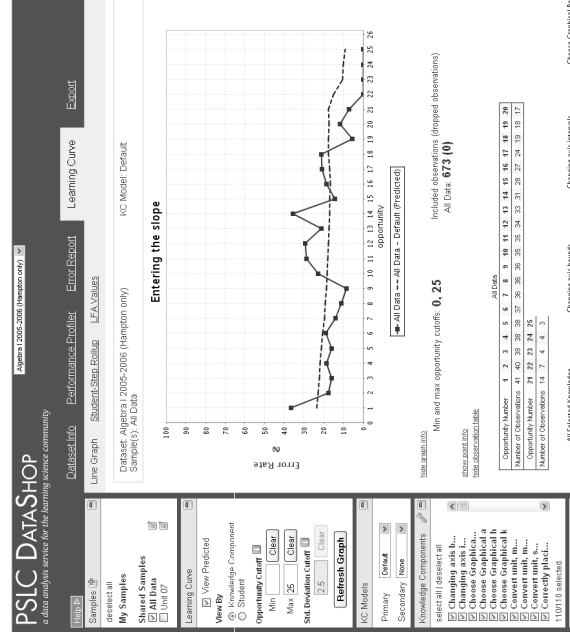
Observation 5 for coder RSB: Clip 143856^o

Time 0:0:	Input: total number of customers on this route
Cell or Context: RTC2	Assessment: RIGHT
Production: UNKNOWN or GIVEN	
Time 20:0:	Input: minutes
Cell or Context: R2C1	Assessment: BUG
Production: Identifying units	
Time 24:0:	Input: hour
Cell or Context: R2C1	Assessment: BUG
Production: Identifying units	
Time 31:0:	Input: years
Cell or Context: R2C1	Assessment: BUG
Production: Identifying units	
Time 45:0:	Input: week
Cell or Context: R2C1	

One excellent source of data

- Led to
 - 7 of 17 full papers at EDM2008
 - 8 of 20 full papers at EDM2009
- Used in exploratory data analysis as well as EDM

<https://pslicdatashop.web.cmu.edu/index.jsp>



PSLC DataShop

- Free access, free to use these data sets in any analysis or publication you want
- Mostly collected in real classrooms by students using educational software for weeks or months or whole years
 - Intelligent tutors, virtual labs, online courseware, optimized drill, ...
 - Math, physics, chemistry, Chinese, ESL, French, domain-general invention & meta-cognition skills
 - ...

DataShop – Dataset Tabs

The screenshot shows the 'Other Datasets' tab in the PSIC DATASHOP interface. It displays a table with columns for 'Dataset', 'Level/Lab', 'Dates', 'Principal Investigator', and 'Status'. The table lists several datasets, including 'Chemistry: Buffer-Solubility_2006' and 'Effects of Collaboration in High School Chemistry Laboratories'. Three callout boxes provide additional information: one states that datasets by a project member or PI can be viewed or edited; another notes that public datasets are viewable by anyone; and a third indicates that private datasets require an email request for access.

Dataset	Level/Lab	Dates	Principal Investigator	Status
Chemistry: Buffer-Solubility_2006	chemistry	Apr 19, 2006 - May 19, 2006	Joel Dawport	complete
Effects of Collaboration in High School Chemistry Laboratories	chemistry	Apr 5, 2007 - Apr 21, 2007	Joel Dawport	complete
Chemistry: Buffer-Solubility_2006	chemistry	Apr 19, 2006 - May 19, 2006	Joel Dawport	complete
Effects of Collaboration in High School Chemistry Laboratories	chemistry	Apr 5, 2007 - Apr 21, 2007	Joel Dawport	complete
Chemistry: Buffer-Solubility_2006	chemistry	Apr 19, 2006 - May 19, 2006	Joel Dawport	complete
Effects of Collaboration in High School Chemistry Laboratories	chemistry	Apr 5, 2007 - Apr 21, 2007	Joel Dawport	complete
Chemistry: Buffer-Solubility_2006	chemistry	Apr 19, 2006 - May 19, 2006	Joel Dawport	complete
Effects of Collaboration in High School Chemistry Laboratories	chemistry	Apr 5, 2007 - Apr 21, 2007	Joel Dawport	complete

Wide range of Exploratory Data Analysis functionality

- A few quick examples

Dataset Info

Dataset Info | Learning Curve | Error Report | Performance Profiler | Export

Overview | Papers and Files | Dataset: Chinese Vocabulary Transfer Lab Study Spring 2006

Dataset Name: Chinese Vocabulary Transfer Lab Study Spring 2006
Project: Knowledge Tracking
Principal Investigator: ppsk

Contributors: LAB Study
Dates: May 24, 2006 - Jul 13, 2006
Location: chrcs
Task: FACT System

Description: Experiment to test the knowledge component hypothesis as it applies to acquisition learned declarative chunks. Over the course of approximately 480 practice items, students were directed to test several transfer hypotheses related to a hierarchical list of all problems for which there is data.

Meta data for given dataset
 • Often includes demographic data

Papers and Files storage
 Published papers using this data set: pre/post-tests, etc.

Problem Breakdown

#	Problem Hierarchy	Problem Name	Step Name
1	UNIT experiment	above_AE4N8B%G4E9%82%8A	dm1P-1
2	UNIT experiment	above_AE4N8B%G4E9%82%8A	dm1P-2
3	UNIT experiment	above_AE4N8B%G4E9%82%8A	dm1P-3
4	UNIT experiment	above_shangqiant	dm1P-4
5	UNIT experiment	above_shangqiant	dm1P-5
6	UNIT experiment	above_shangqiant	dm1P-6
7	UNIT experiment	above_shangqiant	dm1P-7
8	UNIT experiment	above_yw%K7F8%ZFhangqiant wv	dm1P-8
9	UNIT experiment	above_yw%K7F8%ZFhangqiant wv	dm1P-9
10	UNIT experiment	above_yw%K7F8%ZFhangqiant wv	dm1P-10

Dataset Metrics

Number of Students	72
Number of Unique Steps	17
Total Number of Steps	18
Total Number of Transactions	19
Knowledge Component Mentality	0
Pro	0
Res	0
Score	0

Problem Breakdown table

When you'd like to analyze the data, you can use the Problem Breakdown table to access a customizable report.

Performance Profiler

Performance Profiler | Overview | Error Report | Performance Profiler | Export

View measures of Error Rate, Assistance Score, Avg # Hints, Avg # Incorrect, Residual Error Rate

Aggregate by: Step, Problem, KC, Dataset Level

25 [SkillRule: Isolate positive: x=ab, positive] (Hints) %

Knowledge Component	Error Rate (%)
4. Distiller: Consolidate...	~100
5. Distiller: Select Cemb...	~100
6. Distiller: Address...	~100
7. Distiller: Make var...	~100
8. Distiller: Knowledge Component	~100
9. Distiller: Eliminate P...	~100
10. Distiller: Eliminate P...	~100
11. Distiller: Eliminate P...	~100
12. Distiller: Eliminate P...	~100
13. Distiller: Eliminate P...	~100
14. Distiller: Eliminate P...	~100
15. Distiller: Eliminate P...	~100
16. Distiller: Eliminate P...	~100
17. Distiller: Eliminate P...	~100
18. Distiller: Eliminate P...	~100
19. Distiller: Eliminate P...	~100
20. Distiller: Eliminate P...	~100
21. Distiller: Eliminate P...	~100
22. Distiller: Eliminate P...	~100
23. Distiller: Eliminate P...	~100
24. Distiller: Eliminate P...	~100
25. Distiller: Eliminate P...	~100
26. Distiller: Eliminate P...	~100
27. Distiller: Eliminate P...	~100
28. Distiller: Eliminate P...	~100
29. Distiller: Eliminate P...	~100
30. Distiller: Eliminate P...	~100

Multipurpose tool to help identify areas that are too hard or easy

25 [SkillRule: Isolate positive: x=ab, positive] (Hints) %

# students	# steps	# knowledge components
1	1	1

Learning Curve

Learning Curve | Dataset Info | Step Breakdown Table | Error Report | Performance Profiler | Export

Dataset: Chinese Vocabulary Transfer Lab Study Spring 2006
 Samples: All Data

View by KC or Student, Assistance Score, Error Rate, Latency

Visualizes changes in student performance over time

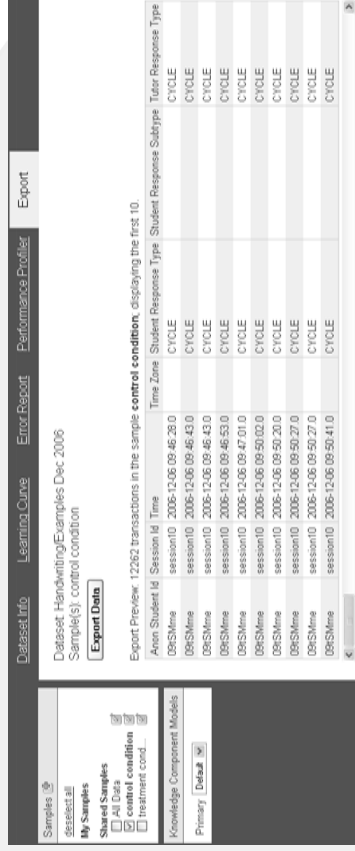
Time is represented on the x-axis as opportunity, or the # of times a student (or students) had an opportunity to demonstrate a KC

All Selected Knowledge Components

Opportunity Number	Knowledge Component	Error Rate (%)
1	cluster_all	~100
2	cluster_all	~100
3	cluster_all	~100
4	cluster_all	~100

Export

- Two types of export available
 - By Transaction
 - By Step
- Anonymized, tab-delimited file
- Easy to import into Excel, Weka, R
- Very fast even for large data sets



Dataset Info Learning Curve Error Report Performance Profile Export

Dataset: HandwritingExamples Dec-2006
Sample(s): control condition

Export Data

Export Preview: 1200 transactions in the sample control condition, displaying the first 10.

Anon Student Id	Session Id	Time	Time Zone	Student Response Type	Student Response Subtype	Tutor Response Type
BRSMne	session10_2006-12-06 09:46:26.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:46:43.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:46:43.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:46:53.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:47:01.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:50:02.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:50:20.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:50:27.0			CYCLE		CYCLE
BRSMne	session10_2006-12-06 09:50:41.0			CYCLE		CYCLE

Plus

- We also host non-PSLC data
 - You have control over who gets access to your data
 - Eliminates storage and access-control problems
 - Makes it easy for EDM researchers looking for data to find your data and use it (and write papers with you or cite your papers)
- If you're interested in storing your data with us, we can chat about this while you're here