Educational Data Mining & Learning Analytics for All: Potential, Dangers, Challenges

Mykola Pechenizkiy

Learning Analytics Seminar
August 30-31, 2011
Utrecht, the Netherlands

Short CV of the Presenter

Mykola Pechenizkiy

*Assistant Professor* at Dept. of Computer Science, TU/e

*Research interests:* data mining and knowledge discovery; Particularly *predictive analytics* for information systems serving industry, commerce, medicine and education.

http://www.win.tue.nl/~mpechen/ - projects, pubs, talks etc.

Major recent EDM-related activities:

- 4th International Conference on Educational Data Mining
  - Mykola Pechenizkiy & Toon Calders (Conference Chairs)
  - Cristina Conati & Sebastian Ventura (PC Chairs)
  - Cristobal Romero & John Stamper (Posters Chairs)

*Handbook of Educational Data Mining*
Motivation for This Talk

• Educational Data Mining (EDM)/Learning Analytics (LA) took off – and the question is – what can or should we do about it?
  – more and more educational data becomes available
  – different kinds of data
  – different kinds of data sources

• Unawareness of many stakeholders of what is already available or what is (potentially) possible with EDM/LA technology
More ICT – More Data Sources
What Kinds of Data We (Can) Have

• Administrative data
  – Who follows which program, who takes which course, registers for an (interim) exam, reexams
  – Demographics, school grades, etc

• Resource usage data
  – Videocollege, owinfo, studyweb, library resources, ...

• LMS (Sakai, Blackboard, Moodle) data
  – More detailed resource usage data
  – Assessment data (online tests)
  – Assignements (text, notes, source code)
  – Forums, collaboration, feedback/help requests
  – Students’ evaluation of learning resources

• Educational games, professional learning, e-Health ...
Objectives of This Talk

• Share my vision on and experience in EDM/LA
• Convince you that
  – EDM/LA is a great thing to do
  – There is technology and plenty of concrete techniques already available for developing and integration bits and pieces of EDM/LA into the education at different levels
  – (If there is time left) There are even more challenging research topics that should be studied for having further success
Outline

• **What EDM/LA is about?**
  – Landscape of tasks & applications
  – Landscape of techniques
  – “Data trumps intuition”
  – Data mining and process mining perspectives

• **Outlook (mostly for the discussion):**
  – promising directions for immediate development and deployment into the educational practice
  – interesting directions for further research
  – Lots of organizational issues to put EDM/LA into everyday’s practice
Can You See the Pattern?

11000000010000000100011001000000110000000000000000001000000
11000000100000001100000000000101000000000010000000
110000010000000000000011000000000000000000010000000
01011100000011011101110010011100111111100001000000000
00011100000011010010110000011001101011111100001000000000
1110000001000000000000110000100000110001000000000000100000
1100000001000000000001100000000000000000010000000000
111101011100000001111000110011000111100011111000001000000000
000001110000000110111100011110011111100011111100110011000000
00000011000000011000000000000000010110000100000000000000000
00000001110111101100000001100011000011000110001100000000000
0000000011101111011111110000111111001111111001110011000000
000000010000000110000001111100000000000000000010110000100000
000000011101111011100000011100011000001100011000000000000
000001110111101111111100001111111001111111001110011000000
0000001111001111011000000011100011000111100011111000010000000
0000001111011110111111100001111111001111111001110011000000
00000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
00000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
000000001110111101111111100001111111001111111001110011000000
Can You See the Pattern?
This Toy Example Illustrates Two Major Problems

• It is difficult for us to see patterns even when the data is homogeneous and “simple”, and the patterns are evidently strong

• It is difficult not to fall into finding/seeing the “patterns” that are really not there
  – Not only our eyes but also tools we use can fool us
KDD/DM/EDM/LA – you name it

• “the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.”  (Hand, Mannila, Smyth)
  – Huge sets of data are being collected and stored
  – Analyzing all data “manually” becomes impossible

• Educational Data Mining can be seen as
  – the process of discovering useful information from the large amount of electronic data collected by educational systems for
    • Learners/students, Teachers, Tutors, Study Advisors, Directors of Education, Educational Researchers, ...
EDM in a Nutshell

Educational data mining: A survey from 1995 to 2005
C. Romero & S. Ventura, Expert Systems with Application 33(1)
What Fields Lay Foundations for EDM/LA

• Knowledge Discovery from Databases or Data Mining, Process Mining
• Information Visualization, Visual Analytics
• Recommender Systems, Search and Information Retrieval
• Social Network Analysis (SNA), Text Mining, Sentiment Analysis
• AI in Education (AIED), Intelligent Tutoring Systems (ITS), Adaptive Educational Hypermedia (AH), User Modeling (UM), Technology Enhanced Learning (TEL),
• Psychometrics, Educational Research
Student Modeling in ITS/AEH

System

User Data

User Model (UM)

Adaptation

Adaptation Effect

get

execute

execute

Learning Analytics Seminar, August 30-31, Utrecht, NL

Educational Data Mining & Learning Analytics for All: Potential, Dangers, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
# EDM: Data ↔ Approach ↔ Knowledge

<table>
<thead>
<tr>
<th>Interactions data</th>
<th>Administrative data</th>
<th>“Feedback” data</th>
<th>Descriptive data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage logs &amp; contexts</td>
<td>Enrolments</td>
<td>Opinions</td>
<td>Demographics</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>Preferences</td>
<td>Characteristics</td>
</tr>
<tr>
<td></td>
<td>Payments</td>
<td>Needs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td><strong>Clustering</strong></td>
<td><strong>Association Analysis, Sequence mining</strong></td>
<td><strong>Process mining</strong></td>
</tr>
<tr>
<td><strong>Categorizing students</strong></td>
<td></td>
<td><strong>Find courses taken together or Popular (parts of) study programs</strong></td>
<td><strong>Understanding study curricular</strong></td>
</tr>
<tr>
<td><strong>Grouping similar students</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Goals
- Identify high risk students
- Predict new student application rates
- Predict students retention/dropout
- Course planning & scheduling
- Faculty teaching load estimation
- Predict demand for resources (library, cafeteria, housing)
- Predict alumni donation

---

Learning Analytics Seminar, August 30-31, Utrecht, NL

*Educational Data Mining & Learning Analytics for All: Potential, Dangers, Challenges*

Mykola Pechenizkiy, Eindhoven University of Technology
Categorizing Students

Student roles in a group project

Learning style dimensions

Learning Analytics Seminar, August 30-31, Utrecht, NL

Educational Data Mining & Learning Analytics for All: Potential, Dangers, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Predicting Student’s Success

- Predicting the group project outcome
- Predicting whether (and when) a student will eventually graduate
- Predicting student drop out
- Discovering exceptional/atypical behavior
Predicting Freshmen Student Dropout at EE Department in TU/e

Learning Analytics Seminar, August 30-31, Utrecht, NL

Eduational Data Mining & Learning Analytics for All: Potential, Dangers, Challenges
Mykola Pechenizkiy, Eindhoven University of Technology
Results of the Case Study

• ~ 80% accurate prediction is possible with very simple decision tree models
Which Questions (tasks, courses, etc) Are Difficult/Easy for Which Students?
# (Visual) Data Analytics in EDM

## Class Seniors 07

<table>
<thead>
<tr>
<th>Student</th>
<th>Absences</th>
<th>Absences per week</th>
<th>...unlearned</th>
<th>...unhandled</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alexander Student</strong></td>
<td>8 9 10 11 12 13 14 15</td>
<td></td>
<td></td>
<td></td>
<td>CORI</td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu</td>
<td>14.4.2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wed</td>
<td>6.4.2011</td>
<td>CORI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu</td>
<td>24.3.2011</td>
<td>CORI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

| **Alma Student** | 8 9 10 11 12 13 14 15 | | | | CORI |
| Date           |          |                   |              |             |       |
| Wed           | 26.1.2011 |                  |              |             | CORI  |
| Total         |           |                   |              |             | 1     |

| **Beatrice Student** | 8 9 10 11 12 13 14 15 | | | | CORI |
| Date           |          |                   |              |             |       |
| Thu           | 28.4.2011 | CORI               | CORI         | CORI        | CORI  |
| Thu           |           |                   |              |             | CORI  |
| Total         |           |                   |              |             | 6     |

| **Beth Student** | 8 9 10 11 12 13 14 15 | | | | CORI |
| Date           |          |                   |              |             |       |
| Wed           | 11.5.2011 |                  |              |             | CORI  |
| Total         |           |                   |              |             | 1     |
• Many more examples of Visual Analytics from MagnaView
Association Analysis/Sequence Mining

- **Given**: a database of sequences
- **Task**: Find all subsequences with support $\geq \text{minsup}$
- **E.g. Administrative data as Sequence Database**
  - **Sequence** is the history of enrolments into courses of a student
  - **Element** (Transaction) is a set of courses taken in a particular semester (or block)
  - **Event** (Item) – one particular course
Application Scenarios

- **Scenario 1**: Find most common types of behavior (and cluster them)
- **Scenario 2**: Find emerging patterns: such patterns, which capture significant
  - differences in behavior of students who graduated vs. those students who did not
  - changes in behaviour of students from year 2006-07 to 2007-08.
  - in both cases we search for such patterns which supports increase significantly from one dataset to another (i.e. in space in the first case and in time in the second case)
- **Scenario 3**: After finding a bottleneck, find frequent patterns that describe it, i.e. for which students it is the bottleneck and why

<table>
<thead>
<tr>
<th>Student</th>
<th>Timestamp</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>S1</td>
<td>2, 3, 5</td>
</tr>
<tr>
<td>A</td>
<td>S2</td>
<td>6, 1</td>
</tr>
<tr>
<td>A</td>
<td>S3</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>S1</td>
<td>4, 5, 6</td>
</tr>
<tr>
<td>B</td>
<td>S3</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>S4</td>
<td>7, 8, 1, 2</td>
</tr>
<tr>
<td>B</td>
<td>S5</td>
<td>1, 6</td>
</tr>
<tr>
<td>C</td>
<td>S1</td>
<td>1, 8, 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student</th>
<th>Graduated</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>B</td>
<td>No</td>
</tr>
<tr>
<td>C</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Common Dangers with Mining

• Implication ≠ causality
  – “Diet Coke → Obesity” or “Intensive Care → Death”

• Forgetting about silent evidence and other biases
  – Data driven intelligence is based primarily on the secondary data analysis, i.e. the data was collected for something else rather than particular hypothesis testing

• Data dredging
  – “Torturing the data until they confess”
  – Overfitting – treading noise or some random behaviours observed in the data as significant patterns

• Discrimination, False discoveries, Interpretability, Redundancy and volume of output knowledge, …
## Simpson’s Paradox

- Success rate of simple and complex operations in two hospitals

<table>
<thead>
<tr>
<th></th>
<th>Academic</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>95%</td>
<td>92%</td>
</tr>
<tr>
<td>Complex</td>
<td>75%</td>
<td>60%</td>
</tr>
<tr>
<td>Total</td>
<td>78%</td>
<td>89%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Academic</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>190/200</td>
<td>920/1000</td>
</tr>
<tr>
<td>Complex</td>
<td>750/1000</td>
<td>60/100</td>
</tr>
<tr>
<td>Total</td>
<td>940/1200</td>
<td>980/1100</td>
</tr>
</tbody>
</table>
Many “mining” Research Challenges

• There are some EDM success stories, but also there are plenty of research questions to be solved, like
  – Use of background knowledge in mining
  – Mining for complex patterns, graph mining/process mining
  – Discovery of emerging and evolving patterns
  – Providing intuitive visualization and explanations
  – …
(Educational) **Process Mining Framework**

- **"world"**
  - exams
  - courses
  - teachers
  - lectures

- models
  - analyzes

- **process models**

- **educational information system**
  - supports/controls
  - specifies
  - configures
  - implements
  - analyzes

- records
  - events, e.g., messages, transactions, etc.

- **process mining**
  - discovery
  - conformance
  - extension
Not to be Mixed with KDD as a Process

Process Analysis/Conformance Checking ex.
Process Discovery Example

Heuristic nets of strict order (left) and flexible order tests (right)
The Objective is to Mine ... 

• structured, easy to understand process models – just like this one

• but ...
... in Reality They May Look Like Spaghetti
Process Model from the Original Log
Process Models from the Clusters

### Diagnosis process

- GYN. ASRT. K. O. SYNTHROLOGIE HEW/HEK
- 1E CONSULT SYNTHROLOGIE HEW/HEK
- ECHO EUP. MEN. FIBRINOS. V. DER. ASR.
- LAST. ASRT. LAB. (complete)
- GYN. JANH. K. O. FIBRINOS. V. DER. ASR.
- GYN. KOMT. K. O. KETAB. V. DER. ASR.
- GYN. KOMT. K. O. KETAB. V. DER. ASR.
- APHROLOG. LABORATORIUM (complete)
- 1E CONSULT KETAB. V. DER. ASR.
- Hysteroscopy (complete)
- Radiotherapy (complete)
- Radiotherapy (complete)

### Treatment process

- GYN. ASRT. K. O. SYNTHROLOGIE HEW/HEK
- 1E CONSULT SYNTHROLOGIE HEW/HEK
- ECHO EUP. MEN. FIBRINOS. V. DER. ASR.
- LAST. ASRT. LAB. (complete)
- GYN. JANH. K. O. FIBRINOS. V. DER. ASR.
- GYN. KOMT. K. O. KETAB. V. DER. ASR.
- GYN. KOMT. K. O. KETAB. V. DER. ASR.
- APHROLOG. LABORATORIUM (complete)
- 1E CONSULT KETAB. V. DER. ASR.
- Hysteroscopy (complete)
- Radiotherapy (complete)
- Radiotherapy (complete)
A More Sound Approach

Isolate a set of standard curriculum patterns and based on this patterns

1. mine the curriculum as an executable quantified formal model and analyze it, or

2. (first) manually devise a formal model of the assumed curriculum and test it against the data.
Example 2-out-of-3 Pattern Check

• At least 2 courses from \{ 2Y420, 2F725, 2IH20 \} must be taken before graduation.
Four Major Types of Learning & Types of Questions EDM Can Assist with

How to (re)organize the classes, or assessment, or placement of materials based on usage and performance data.

How to identify those who would benefit from provided feedback, study advice or other help; How to decide which kind of help would be most effective?

How to help learners in (re-)finding useful material, done whether individually or collaboratively with peers.

How to help learners in (re-)finding useful material, done whether individually or collaboratively with peers.
A Wide Scope of “learning”

• Traditional education at primary, secondary and high-school (algebra tutors), and University levels
• eLearning and Blended learning (LMS like Sakai, Moodle, Blackboard; SQL tutor)
• Professional education – (pilot, military simulators)
• Rehabilitation, elementary skills like reading and performing arithmetic operations (“Neure”, “Ekapeli”)
• eHealth, Patient education (Philips “Motiva”, RPM)
• Learning becomes more informal, mobile, social and ICT/data/information intense in all areas
  – EDM is just making the first steps to address this
EDM Evolution

• **Phase I**: Educational research curiosity driven
  – Come up with a hypothesis, collect data, test and publish the results

• **Phase II**: DM research curiosity driven
  – Data is there, collected because it can be collected
  – Do EDM, generate interesting hypothesis, test, and publish the results

• **Phase III**: Educational needs driven
  – Technology is there, knowledge is there,
  – Time to start the valorization - Carnegie Learning Lab Success

• **Phase IV**: Synergy of R&D
  – Understanding that not all know-how is there
  – Understanding (by researchers) that assessment in non-lab settings is crucial
  – Understanding (by stakeholders) that R&D forms a natural cycle
From EDM’s past and presence to EDM of tomorrow

• Lyrics by Jack Mostow, CMU
  – All my data from my tutor, I put on a USB
    That I accidentally swallowed
    Now my study's history.
  – All my data,
    Educational data that I mine,
    Now is lost and gone forever --
    Dreadful sorry, data mine
... full text can be found at EDM 2011 website

• Indeed, this relates primarily to Phases I & II, while we need to think more about Phases III & IV
  – how to organize data collection to turn data instances into “first-class citizens” rather than some byproduct
  – how to address ethical, privacy and scrutability issues
  – how to enable and organize EDM activities at different levels, individual, institutional, national levels

Reflects on not easy early days of EDMers
Take Home (& for the discussion) Messages

• Many stakeholders: students, lecturers, tutors, study advisors, directors of education, educational institution and national level

• Lot’s of potential benefits for each category of stakeholders

• Popular data mining problem formulations fit well to the educational domain and there are state-of-the-art techniques to address them

• Promising directions for further research from the applications perspective
Still Open (and often not technology related) Questions

• Privacy and ethics
  – What is EDMer’s philosophy?
  – Is EDM always ethical?
  – Is EDM a threat to privacy?
• Students, educators, directors

• Who are EDM stakeholders?
• Organization of data collection processes
• Organization of EDM activities
Acknowledgements

• Thanks to many colleagues at TU/e:
  – Wil van der Aalst
  – Nikola Trcka
  – Ekaterina Vasilyeva
  – Toon Calders
  – Paul De Bra
  – Jan Vleeshouwers
• SURFfoundation
  – John Doove
• ...
• I wish I could acknowledge funding agencies here
Thank you!

• Questions
• Suggestions
• Collaboration

• all warmly welcome