Educational Data Mining: Preliminary results at University of Porto

Pedro Strecht
João Mendes Moreira
Carlos Soares

EUNIS BUSINESS INTELLIGENCE CONFERENCE
MARCH 6TH-7TH 2014
PARIS
Summary

• Data Analysis in Education
• ... at the University of Porto
• An illustrative example of an EDM task
• Conclusions and Future work
Data Analysis in Education

- For a few decades higher education institutions manage their data using University Information Systems (UIS).

- The growing adoption of UIS allowed research to move towards automatic knowledge discovery from academic databases.

- Over the past 10 years there has been an increase on research using data mining techniques to discover phenomena in the data.

- An example of application of data mining is:
  - Predicting the success or failure of student enrolled in a course.
  - Learning the reasons behind it.
University of Porto

- Founded in 1911
- 14 faculties, 1 business school
- ~700 study programs
- ~32 000 students, ~2 000 teachers and researchers, ~1 800 administrative staff
- University Information Systems began being developed in-house and explored since 1992
- The SIGARRA system had a major improvement in 2012 which prompts the University to improve their processes using BI and DM
Data Analysis in Education

Educational big data

- Academic information
  - Administrative data
  - Pedagogical data

- Teaching and learning environments

- Others sources
  - Other types of data
Data Analysis in Education
Overview of research

Educational data

Administrative data

Pedagogical data

Other types of data

Focus

Automated discovery

Educational Data Mining

Statistics

Data mining

Machine learning

Learning Analytics

Information science

Sociology

Psychology

Interpretation of data

Main goals

• Inform education practice
• Discover phenomena in data

• Influence education practice
• Adapt teaching to students
Educational DM & Learning Analytics at U.Porto: general perspective

**Predictive analysis** (aka EDM)

**Descriptive analysis** (aka LA)
- Social Network Analysis
- Cluster analysis
- Statistical data analysis
- Exploratory data analysis
- Business Intelligence

**Data Warehouse**

**Improve Processes**
- Thwart attrition
- Adapt student tutoring
- Identify research collaboration opportunities
- Adapt management workflow

**Feature engineering**

**Educational data**
- Administrative data
- Pedagogical data
- Other types of data

**Techniques**
- Classification
- Regression
- Anomaly detection
- Pattern mining
- Text mining

**Processes**
- Improve processes
- Thwart attrition
- Adapt student tutoring
- Identify research collaboration opportunities
- Adapt management workflow

**Outcomes**
- Improve processes
- Thwart attrition
- Adapt student tutoring
- Identify research collaboration opportunities
- Adapt management workflow

**Strategies**
- Improve processes
- Thwart attrition
- Adapt student tutoring
- Identify research collaboration opportunities
- Adapt management workflow
Learning Analytics at U.Porto: current work

Descriptive analysis (aka LA)

Front-ends such as
• Pivot tables
• Report tools

OLAP cube
Education

OLAP cube
Finance

Data Warehouse

Educational data

Administrative data

Pedagogical data
Educational DM at U.Porto: current work

Improve Processes

- Discovery of students’ attrition factors
- Comparative study on differences in factors of attrition
- Development of strategies to thwart attrition

Descriptive analysis (aka LA) → Feature engineering

Predictive analysis (aka EDM)

- Student attrition models
- Pedro Strecht (Ph.D. thesis)
  - Curriculum mining
  - Social Network Mining

- Student attrition models
- Luís Cruz (Ph.D. thesis)
  - Curriculum mining
  - Living Analytics

Data Warehouse

Educational data

- Administrative data
- Pedagogical data
- Other types of data
Summary

- Data Analysis in Education
- ... at the University of Porto
- An illustrative example of an EDM task
- Conclusions and Future work
An illustrative example of an EDM task

- System to predict if a student will pass or fail a course
- Using administrative data from UIS
- Three different processes

Diagram:

1. Data extraction
2. Classifiers training and model analysis
3. Classifiers evaluation

Variables’ importance measures
Models’ performance measures
Data extraction

- 14 variables extracted relating to each student

<table>
<thead>
<tr>
<th>Group</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographic information</td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
</tr>
<tr>
<td></td>
<td>Marital status</td>
</tr>
<tr>
<td></td>
<td>Nationality</td>
</tr>
<tr>
<td></td>
<td>Displaced</td>
</tr>
<tr>
<td></td>
<td>Scholarship</td>
</tr>
<tr>
<td></td>
<td>Special needs</td>
</tr>
<tr>
<td>Admission information</td>
<td>Type of admission</td>
</tr>
<tr>
<td>Enrollment information</td>
<td>Type of student</td>
</tr>
<tr>
<td></td>
<td>Status of student</td>
</tr>
<tr>
<td></td>
<td>Years of enrollment</td>
</tr>
<tr>
<td></td>
<td>Delayed courses</td>
</tr>
<tr>
<td></td>
<td>Type of dedication</td>
</tr>
<tr>
<td>Financial information</td>
<td>Debt situation</td>
</tr>
</tbody>
</table>

- 8 courses were selected
Data extraction

- Data set sample for course Mathematics II

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Marital status</th>
<th>Nationality</th>
<th>Displaced</th>
<th>Scholarship</th>
<th>Special needs</th>
<th>Type of admission</th>
<th>Type of student</th>
<th>Status of student</th>
<th>Years of enrollment</th>
<th>Delayed courses</th>
<th>Type of dedication</th>
<th>Debt situation</th>
<th>Approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>m</td>
<td>s</td>
<td>pt</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>r</td>
<td>r</td>
<td>o</td>
<td>0</td>
<td>0</td>
<td>f</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>32</td>
<td>m</td>
<td>m</td>
<td>pt</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>tcs</td>
<td>r</td>
<td>o</td>
<td>8</td>
<td>12</td>
<td>p</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>18</td>
<td>f</td>
<td>s</td>
<td>pt</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>r</td>
<td>r</td>
<td>o</td>
<td>0</td>
<td>0</td>
<td>f</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>18</td>
<td>m</td>
<td>s</td>
<td>pt</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>r</td>
<td>r</td>
<td>o</td>
<td>0</td>
<td>0</td>
<td>f</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>22</td>
<td>m</td>
<td>s</td>
<td>br</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>to</td>
<td>r</td>
<td>o</td>
<td>1</td>
<td>0</td>
<td>f</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>
Classifiers training and model analysis

Experimental setup

1. Data extraction
   - Academic database

2. Classifiers training and model analysis
   - Course data set
   - Classifier training
   - Decision tree model
   - Model analysis
   - Variables' importance measures

3. Classifiers evaluation
   - Courses data set
   - Models’ performance measures

4. Variables’ importance measures
Classifiers predict categorical class labels

Students are classified as either having as either having passed or failed

Example of decision tree for course Mathematics II:
Classifiers training and model analysis
Preliminary results

- Variables’ importance measure for each course

<table>
<thead>
<tr>
<th>Course</th>
<th>#P</th>
<th>Age</th>
<th>Sex</th>
<th>Marital status</th>
<th>Nationality</th>
<th>Displaced</th>
<th>Scholarship</th>
<th>Special needs</th>
<th>Type of admission</th>
<th>Type of student</th>
<th>Status of student</th>
<th>Years of enrollment</th>
<th>Delayed courses</th>
<th>Type of dedication</th>
<th>Debt situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic History</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organic Chemistry II</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroanatomy</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomy I</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomy II</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics II</td>
<td>4</td>
<td>76.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introduction to Linear Signals and Systems</td>
<td>3</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classifiers evaluation

Experimental setup

1. Data extraction

1. Academic database

Courses data set

2. Classifiers training and model analysis

3. Classifiers evaluation

- Performance measures 1
- Performance measures 2
- Model performance measure

Classifiers training and model analysis

- Variable importance measures
- Models' performance measures

Variables' importance measures

Course data set

10-fold partitioning

- Evaluation 1
- Evaluation 2
- Evaluation 10

Stratified 10-fold cross validation

10-fold partitioning

- Train data
- Test data

For each course

Data extraction

Academic database

Classifiers evaluation

Course data set

10-fold partitioning

Classifiers training

Decision tree model

Performance measures

Averaging
Classifiers evaluation

Performance results

- Model performance for each course (10 experiments)

<table>
<thead>
<tr>
<th>Course</th>
<th>Number of examples</th>
<th>Category distribution (%)</th>
<th>F1 (avg ± std.dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic History</td>
<td>656</td>
<td>72 28</td>
<td>0.83 ± 0.003</td>
</tr>
<tr>
<td>Organic Chemistry II</td>
<td>562</td>
<td>21 79</td>
<td>0.10 ± 0.030</td>
</tr>
<tr>
<td>Neuroanatomy</td>
<td>542</td>
<td>94 6</td>
<td>0.96 ± 0.001</td>
</tr>
<tr>
<td>Marketing</td>
<td>519</td>
<td>90 10</td>
<td>0.95 ± 0.002</td>
</tr>
<tr>
<td>Anatomy I</td>
<td>518</td>
<td>73 27</td>
<td>0.85 ± 0.003</td>
</tr>
<tr>
<td>Anatomy II</td>
<td>477</td>
<td>73 27</td>
<td>0.84 ± 0.004</td>
</tr>
<tr>
<td>Mathematics II</td>
<td>476</td>
<td>61 39</td>
<td>0.78 ± 0.005</td>
</tr>
<tr>
<td>Introduction to Linear Signals and Systems</td>
<td>475</td>
<td>55 45</td>
<td>0.71 ± 0.099</td>
</tr>
</tbody>
</table>
Conclusions

- There is a global effort of University of Porto to improve their processes using BI and DM

- This work presents the preliminary experiments on Educational Data Mining
  - Using administrative data
  - Collecting 14 variables from students enrolled in 8 courses
  - Interpreting results from decision tree models

- Results indicate that
  - Decision trees are quite different from one another
  - Delayed courses is the most important variable
    - Will this pattern hold if more courses are used?
  - Model performance is quite acceptable overall
Future work

- Study the reasons for the variability of variables in each course

- Alternatives to combine decision trees into
  - a single consensual tree
  - small set of trees

  that represent the general knowledge about the success/failure behavior across all the University

- Although the focus is on EDM, such an approach will be interesting for other areas of application
Questions