Educational Data Mining: Preliminary results at University of Porto

Pedro Strecht
João Mendes Moreira
Carlos Soares
Summary

- University of Porto (brief presentation)
- Data Analysis in Education
- An illustrative example of an EDM task
- Conclusions
- Future work
University of Porto
Brief presentation

- 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 System developed in-house and explored since 1992
  - Evolved into the SIGARRA system in 2003

- Major improvement in 2012 which paved the way to BI and DM initiatives
Data Analysis in Education

Context

- Since the past 10 years there has been an increase on research using data mining techniques to discover phenomena in data
  - Educational Data Mining (EDM)
  - Learning Analytics (LA)

- An example of application of data mining is:
  - Predicting the success or failure of student enrolled in a course
  - Learning the reasons behind it
Data Analysis in Education

Sources of educational data

- Academic information
- Teaching and learning environments
- Others sources
An illustrative example of an EDM task

- Build a system to:
  - Predict if a student will pass or fail a course
  - Understand which are the most important variables used in prediction
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>
Data extraction

- SQL queries to create CSV files
- ~3 hours to collect data for one academic year
**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>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>
<td></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>tcs</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

Classifiers training and model analysis

Academic database ➔ (1) Data extraction ➔ Courses data set ➔ (2) Classifiers training and model analysis ➔ Variables’ importance measures

(3) Classifiers evaluation ➔ Models’ performance measures
Classifiers predict categorical class labels

Students are classified 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>11.5</td>
<td>72.1</td>
<td>100.0</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>100.0</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>36.7</td>
<td>20.9</td>
<td>18.4</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics II</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.4</td>
<td>87.4</td>
<td>79.6</td>
<td>100.0</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>93.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.9</td>
</tr>
</tbody>
</table>
Classifiers evaluation
Experimental setup

1. Data extraction
   - Academic database

2. Classifiers training and model analysis
   - Classifier evaluation
   - Decision tree model
   - Performance measures

3. Classifiers evaluation
   - Models' performance measures

- Course data set
  - 10-fold partitioning
    - stratified 10-fold cross validation
      - for each course

- Classifier training
  - Decision tree model
  - Performance measures

- Model performance measure
## 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</td>
<td>28</td>
</tr>
<tr>
<td>Organic Chemistry II</td>
<td>562</td>
<td>21</td>
<td>79</td>
</tr>
<tr>
<td>Neuroanatomy</td>
<td>542</td>
<td>94</td>
<td>6</td>
</tr>
<tr>
<td>Marketing</td>
<td>519</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>Anatomy I</td>
<td>518</td>
<td>73</td>
<td>27</td>
</tr>
<tr>
<td>Anatomy II</td>
<td>477</td>
<td>73</td>
<td>27</td>
</tr>
<tr>
<td>Mathematics II</td>
<td>476</td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td>Introduction to Linear Signals and Systems</td>
<td>475</td>
<td>55</td>
<td>45</td>
</tr>
</tbody>
</table>
Conclusions

- There is an 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

- Create models for every course in the university
- Study the reasons for the variability of variables in each course
- Study the alternatives to combine decision trees into a small set of trees from clusters of trees
- Although the focus is on EDM, such an approach will be interesting for other areas of application
Future work
Experimental set-up (work in progress)

1. Data extraction
2. Models training and analysis
3. Models evaluation
4. Models clustering
5. Models merging
6. Merged models evaluation

Academic database → Data extraction → Courses data set → Variables’ importance measures → Models data set

Models’ performance measures → Merged models’ performance measures

Merged models data set → Models’ cluster → (4) Models clustering
Questions
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

Main goals
- Inform education practice
- Discover phenomena in data
- Influence education practice
- Adapt teaching to students

Interpretation of data
<table>
<thead>
<tr>
<th>Course</th>
<th>Program</th>
<th>Faculty</th>
<th>Academic year</th>
<th>Semester</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic History</td>
<td>Economics</td>
<td>Economics</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>656</td>
</tr>
<tr>
<td>Organic Chemistry II</td>
<td>Pharmaceutical Sciences</td>
<td>Pharmacy</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>562</td>
</tr>
<tr>
<td>Neuroanatomy</td>
<td>Medicine</td>
<td>Medicine</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>542</td>
</tr>
<tr>
<td>Marketing</td>
<td>Economics</td>
<td>Economics</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>519</td>
</tr>
<tr>
<td>Anatomy I</td>
<td>Medicine</td>
<td>Medicine</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>518</td>
</tr>
<tr>
<td>Anatomy II</td>
<td>Medicine</td>
<td>Medicine</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>477</td>
</tr>
<tr>
<td>Mathematics II</td>
<td>Economics</td>
<td>Economics</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>476</td>
</tr>
<tr>
<td>Introduction to Linear Signals and Systems</td>
<td>Electrical and Computers Engineering</td>
<td>Engineering</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>475</td>
</tr>
</tbody>
</table>
Variable importance is the % of training set samples under a node:

$$I\downarrow p = \frac{\#\text{examples in node}}{\text{total examples}}$$

In course Mathematics II the total examples is 476

\[
I\downarrow \text{delayed courses} = \frac{476}{476} = 100.0\
I\downarrow \text{type of admission} = \frac{416}{476} = 87.4\
I\downarrow \text{years of enrollment} = \frac{379}{476} = 79.6\
I\downarrow \text{age} = \frac{364}{476} = 76.4\
\]
The predictions made by a classifier can be described in a confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>Predicted positive</th>
<th>Predicted negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual positive</td>
<td>$TP$</td>
<td>$FN$</td>
</tr>
<tr>
<td>Actual negative</td>
<td>$FP$</td>
<td>$TN$</td>
</tr>
</tbody>
</table>

The evaluation measures are:

- **Precision:** $P = \frac{TP}{TP+FP}

- **Recall:** $R = \frac{TP}{TP+FN}$

- **F1 measure:** $F1 = \frac{2PR}{P+R}$