Using Data Mining to Evaluate Colorado Public Schools Performance

Aaron Holt, Anas Salamah, Hui Soon Kim

Motivation

- Kaggle competition by Colorado School Grades
  - Use school data supplied by the Colorado Department of Education to visually uncover trends in the public school system
  - Where are the good schools?
  - How have grades changed over time?
  - Which schools are improving?

- Schools want their students to improve!
- Large amounts of data with little analysis
- End goal to use uncovered trends to help educators improve schools

Current Literature

Data Mining Technique

1. Frequent Pattern mining(Apriori, FP-growth)
    - Find frequent itemset given support value-> Find strong association and correlation rule -> Eventually, find the significant attributes to affect the school performance
2. Classification(Naive Bayes, KNN, ID3, etc.)
    - Predict school performance using chosen attributes

External attributes to consider(combining external data set)

1. Family Income
2. Parent’s education level
3. Racial Distribution
4. Single Parent home, etc.

Proposed Work

Initially process and collect data:

- Colorado school data comes in multiple files with different attribute types
- There is missing data for some schools
- Census and crime data come in a different format

Challenges:

- Colorado department of education data lists the school and district names but not location
- Will have to map schools to their respective city and county to include census and crime data

Proposed Work

Frequent pattern and trend analysis:

- Start with static trends:
  - Which attributes correlate to positive academic achievements?
  - Negative achievement?
- Also consider trends over time:
  - How has Colorado education changed?
  - Do previous trends still hold true?
- Explore some of the trends found in the Kaggle competition:
  - Map where the good schools are, which schools are improving etc.
- Focus on new trends from external data
  - Family income, education level, crime data

Proposed Work

Why a classifier?

- Colorado’s population is growing and new schools continue to be built
- Predict how well schools will rank based on given school attributes

Explore various classification techniques

- Decision trees, logistic regression, SVM’s

Practice feature engineering

- Find meaningful features to accurately classify rank
- Previous trend analysis should be useful

Evaluation - Trend Analysis

The Colorado Department of Education ranks schools on the following variables:

Academic Achievement

- Is a school meeting Colorado's model content standards?
  - Primarily tested through state assessments
- Graduation rate

Academic Growth:

- A relative measure of academic achievement
- How much did a student learn compared to similar students

Academic Growth Gaps

- Academic growth for disadvantaged students

Evaluation - Classifier

How well does the classifier predict the specified attribute?

1. Split the data into training and test sets
2. Train on a majority of the data (80%)
3. Test on the remaining data (20%)
4. Evaluate based on percentage of correct classification within a margin
    - The margin will be defined based on the specified attribute

- It is very difficult to determine the upper limit for a given classification task.
- Compare different classification methods to try and find a good solution

Milestones

Project Deliverable	**Expected Completion Date**
Data collection and preprocessing	3-6-2015
Static frequent pattern analysis	3-16-2015
Pattern analysis for trends over time	3-30-2015
Classification task	4-10-2015

Questions?