Using Data Mining to Evaluate Colorado Public Schools Performance

Aaron Holt
University of Colorado Boulder
2900 E College Ave Ap#8
Boulder, CO 80303
+1 (719) 201-0277
aaron.holt@colorado.edu

Anas Salamah
University of Colorado Boulder
4670 White Rock Circle
Boulder, CO 80301
+1 (720) 240-6170
anas.salamah@gmail.com

Hui Soon Kim
University of Colorado Boulder
1855 Athens St Apt. 304
Boulder, CO 80301
+1 (303) 960-6480
huki2996@colorado.edu

ABSTRACT

In this paper, we describe the process we used to apply data mining techniques on Colorado public education data. We want to investigate the performance of public schools in Colorado based on the overall given grade by analyzing the data to find frequent item sets and finding the relationships between these items, as well as, validating the performance based on other outer sources such as US census data which includes financial data the different cities in Colorado. We also want to build a classifier that predicts the grade of a school based on its associated information. Finally, we want to validate our classifier by holding out some data as test data and measuring the accuracy of our classifier.

Keywords

Educational Data Mining, frequent pattern analysis, Association rules, Classification.

1. Motivation

Since the introduction of the No Child Left Behind (NCLB) in 2001, data driven decision-making has become an increasingly important topic for schools nationwide. In order for schools to prove they have met the Adequate Yearly Progress requirements defined in NLCB, schools must collect a significant amount of data pertaining to their student demographics, graduation rate, grades, and state assessment scores [1]. The collected data is used to show whether or not a school has achieved its yearly progress goals and to identify areas of improvement.

One of the many challenges schools face is finding valuable ways to use the large amounts of data given. As is often the case with big data, schools find themselves being “data rich but information poor” [2]. Principal Dr. Gregory Decker of Lead Mine Elementary succinctly stated the problem saying "Receiving test data in July is like driving a school bus looking out the rearview mirror. I can see where my students have been but I cannot see where we are going."[2] Thus schools have been attempting to look at information over time to track student achievement. While there has been progress in this area, data from external sources such as crime data or US census data is often excluded. In order to improve upon the current methodologies, patterns from Colorado school collected data and external sources, both statically and over time should be considered.

The overarching goal of this project is to find meaningful patterns that highlight successes and failures in schools and develop prediction model to evaluate the school performance by combining available public school data of Colorado and other external data such as US census and criminal data. Most schools measure success based on student grades, test scores, and graduation rates. Thus finding correlations between these attributes and those from outside sources is a priority.

The dataset currently proposed comes from several sources. The majority of the data ware collected by the Colorado Department of Education (CDE) and R-Squared Research. The CDE data contains information about students’ grades, test scores, ethnicity, and progress at every Colorado public school. The next source comes from the US census, which includes financial data in different cities, counties and the state as a whole. The final source is crime data from the Colorado Freedom of Information Coalition. The combination of these sources will hopefully yield new and useful correlations for the improvement of Colorado schools.

2. Literature Survey

2.1 Data Mining Technique

This section reviews the different data mining techniques that may be used for finding significant attributes to affect the school performance and predict the overall grade.

Frequent item sets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association and correlation rules. In order to find correlation between independent variables and dependent variable, frequent pattern analysis has to be done at first. The Apriori algorithm and FP-growth algorithm [3] can generate all frequent item sets with a specific support value. And then finding the association

and correlation rules for given frequent item sets, which results in finding the significant attributes that affect the school performance.

Classifications may be used for predictions of school performance. The techniques that are reviewed are Naïve Bayes [4], KNN [4] and ID3 [5]. Bayesian classifier is the supervised data mining technique used to take decisions under uncertain conditions. The concept of probability is used to classify the new entities by It looking into past data to predict the performance of newly entered schools. The value of a given class is conditionally independent of the values of other attributes. In Bayesian analysis the final classification is produced by combining both of the information i.e. the prior and the likelihood, to form the posterior probability using the so-called Bayes. K nearest neighbor algorithm (KNN) is known as lazy learner that makes predictions based on KNN labels assigned to train sample. K nearest neighbor is determined by measuring the distance between the new entered query and previously known samples. The bulk of K nearest neighbors is taken for the prediction of the entered query once the K nearest neighbor is assembled. ID3 is one of the popular DT algorithms that deal with the nominal data sets. ID3 is a classical version of the decision tree induction. It mainly works on the selections of attributes at all the levels of decision tree that are based on entropy. Basically top-down greedy search approach is followed to construct the tree, where each attribute is tested on every tree node.

2.2 External Attributes to affect School Performance

Some external attributes such as family income, parent's education, racial distribution and single parent Etc. need to be considered because they can affect children's school performance. For example, According to the survey, "Current family income has significant effects on a child's math and reading test scores"[6], "Socioeconomic variables do influence parenting beliefs and behaviors and that these parenting variables influence subsequent change in child's achievement" [7], "Although scores have increased for both Black students and White students, on average Black students do not perform as well as their White peers."[8], "Lower high school graduation rates, lower GPAs, and greater risk for drug abuse are some of the negative outcomes associated with growing up in a single-parent home"[9]. So various attributes from other data sources that may affect the school performance will be combined in this project.

3. Proposed Work

3.1 Data Collection and Preprocessing

The initial work involves collecting and processing data. The dataset used in this project consists of Colorado Department of Education (CDE) and R-Squared Research data and other sources such as US census and criminal data in different formats each with varying attribute types. Data preprocessing is an important step in which missing values are filled up, redundancies are removed and filtering is performed so that the database will be ready for use. In order to analyze these data in combined manner, the data integration is performed using common attributes between data sets. According to the attribute types such as normal, numerical, binary and ordinal, a proper data preprocessing technique such as normalization, discretization and linearization are needed.

3.2 Frequent Pattern Analysis

Before moving on to trend analysis over time, it will be useful to establish static trends. There are several genres of trends to search for. The first set involves searching for simple location based items such as where the best and worst schools are located.

Next correlations between various attributes and student achievement can be explored. For example, one could explore how family income of the county affects grades and graduation rates of schools in that county. Results could establish a correlation between wealth and performance.

The discovered static correlations will be used to guide the search for trends over time. Trends over time are likely more valuable as they indicate whether a given strategy is working or not. Similar to the static analysis, trends correlating attributes to student achievement will be explored first.

3.3 Modeling a Classifier

After deciding the significant attributes to influence the school grade, we can model some classifier to predict the school performance using them.

4. How to Evaluate

There are several core evaluation metrics for this project. For both the static and over time analysis, this will primarily be how do various attributes correlate to student achievement. Student achievement for static analysis will explore how an attribute affects overall grades and graduation rate. In the data, the overall grade is an average of reading, writing, math and science. Each school along with its type (Elementary, Middle, or High school) is assigned a grade. Some schools combine more than one type and thus are given a grade for each type. Individual

subject grades will also be considered in trend analysis. The trend over time analysis will be similar except for overall grade is replaced with overall growth where growth is defined as the change in individual subject scores over time. Finally, the classifier will be evaluated based on the percentage of time it can predict school grades and graduation rate within a given margin. In order to evaluate our grade classifier, we need to split our data to training and testing data. Also we need to understand how these grades are calculated.

5. 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

6. References

[1] Mandinach, Ellen, Margaret Honey, and Daniel Light. *A Theoretical Framework for Data-Driven Decision Making*. Rep. San Francisco: AERA, 2006. Print.

[2] Salpeter, Judy. *Data: Mining with a Mission.* Rep. N.p., 15 Mar. 2004. Web. 20 Feb. 2015.

[3] Survey on Frequent Pattern Mining Bart Goethals HIIT Basic Research Unit Department of Computer Science University of Helsinki

[4] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi and M.A. Sid-Ahmed, "Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers", 2007 International Conference on Convergence Information Technology, IEEE DOI 10.1109/ICCIT.2007.148

[5] An Implementation of ID3 --- Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia

[6] The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit- By Gordon B. Dahl and Lance Lochner

[7] How Does Parents' Education Level Influence Parenting and Children's Achievement?- Pamela E. Davis-Kean University of Michigan

[8] Achievement Gaps U.S. Department of Education NCES 2009-455, How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress Statistical Analysis Report

[9] Academic Achievement of Children in Single Parent Homes: A Critical Review-by Mark S. Barajas Western Michigan University