From Data to Actionable Knowledge: A Collaborative Effort with Educators

J. BARRETT, E.B. CARUTHERS, K. GERMAN, E.S. HAMBY, R.M. LOFTHUS, S. SRINIVAS

Xerox Research Center Webster, USA

AND

E. ELLS

Klem South Elementary School (Webster, NY), USA

Teachers, school administrators, and researchers at Xerox have started a collaborative effort to leverage educational data from schools to target the individual needs of students in the classroom. In this paper, we report on our observations of using a data driven methodology to individualize education. First, we discuss the common types of data that schools now collect and the different types of actionable knowledge that could be valuable to students, teachers, principals, district, state and national administrators. We also discuss ways that data mining and analysis can 1) provide guidance for each individual student, 2) suggest temporary clustering of students for targeted instruction, and 3) combine data from multiple tests with demographic and other factors to separate the effects of curriculum and teaching changes from uncontrollable factors that affect student learning. We welcome academic collaborators, who can work with us to find solutions to challenges in mining of educational data.

Key Words and Phrases: Educational Data Mining, Educational Recommendation System, Student Clustering, Pattern Detection

1. INTRODUCTION

Improving the quality and effectiveness of education is an issue of high priority to nations across the globe. In the US, both the general public and the policy makers have shown renewed interest in improving education through programs such as Race to the Top and No Child Left Behind. Along with these policy reforms, there also have been substantial technology enhancements in the classroom. Instead of treating the classroom as a black box, there is serious effort to understand the dynamics and the inner workings of this complex system [Black 98]. Standardized tests, which can be indicators of the strengths and the weaknesses of the students, are being administered in the classroom. Curriculum based formative assessments are growing in popularity among schools. These assessments can be used to gather highly granularly data that describes each student's zone of proximal learning which, in turn, informs instructional changes. Black et al. in their influential publication on the advantages of formative assessment [Black 98] point out that they know of no other way of raising standards for which such a strong prima facie case can be made. Professional Learning Communities [Hord 97] are being developed for the educators to encourage data driven decision making and sharing of best practices.

The combination of the policy and the technology enhancements along with other factors such as global competitiveness and the premonition of an innovation driven economy, have created a tipping point for both the opportunity and the need to individualize education, i.e., customizing the learning experiences according to the individual needs of the students. At Xerox Research, we have an ongoing project that complements the existing paper workflows in the schools and helps them to gather, manage, and store fine grained data on student performance. We also have started

Authors' addresses: Jan Barrett, Edward Caruthers, Eric Hamby, Robert Lofthus, Sharath Srinivas, Xerox Corporation, 800 Phillips Road, Webster, NY Email: Jan.Barrett@xerox.com, Edward.Caruthers@xerox.com, Krinstine.German@xerox.com, Eric.Hamby@xerox.com, Robert.Lofthus@xerox.com, Sharath.Srinivas@xerox.com

Erin Ells, Klem South Elementary School, 1025 Klem Rd, Webster, NY, Email: Erin_Ells@websterschools.org

extensive data mining of student performance on both summative and formative assessments from local schools. We believe that by combining information from multiple sources and by analyzing highly granular performance data, deep insights that can help understand the instructional needs for each student can be extracted.

Many of the challenges facing those involved in educational data mining is also shared by the broader data mining data community [Steinbach 03], e.g., there has been an explosion in the amount of data that is available, data is high dimensional because of integration from multiple sources, and the useful patterns are masked by several layers of noise. The popularity of formative assessments has further exacerbated this problem as data is more granular and is being collected more frequently. A unique aspect of educational data mining is that there is a hierarchy of end users for the actionable knowledge that could be extracted from educational data. At the first level, there are students who could assess their own performance in order to identify their strengths and weaknesses. At the next level, teachers could assess the effectiveness of their instructions and adjust their lesson plans accordingly. The school administrators, who are at the next level, could perform an internal appraisal of the performance of their classes at the grade level and at the same time could compare the performance of their schools with other similar schools. Finally, the district and the state level administrators could detect and proactively fix system-wide problems. Thus, though the underlying data is the same, the analysis and the reporting has to be customized for the users at each level in this hierarchy for it to be truly actionable. Ultimately, the goal of all users of the data is to address the needs of the students. Fig 1 shows the hierarchy of end users for data mining in education and their knowledge requirements.

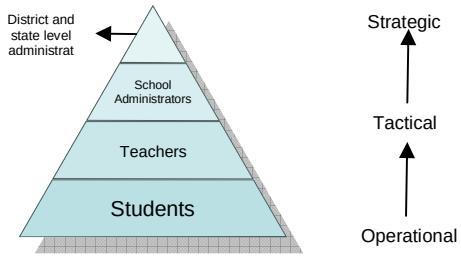

Fig 1: The Education hierarchy and the knowledge requirements

Recently, there has been a lot of interest in the Artificial Intelligence and Machine Learning community to mine data from an educational setting [Aleven 06, Baker 07]. The focus of much of the work in this field has been on cognitive tutors, educational software technology, modeling student learning, and forecasting student performance. The focus of our work is on mining data that is much more closely tied to the school, curriculum and students. The purpose of this paper is two-fold: First, to share our experiences working with students, teachers and administrators in order to identify their problems that could be solved by a data driven methodology, and second, to invite open innovation partners, who would be interested in collaborating with us to find solutions to these problems. The reminder of this article is organized as follows: Section 2 describes the common types of data that schools are currently collecting on student performance and newer varieties they could collect. In Sections 3, we present challenging issues facing students, teachers, parents and administrators. Some preliminary thoughts on data driven methodologies that can be used to tackle these problems are also discussed. The final section summarizes this paper and articulates our next steps.

2. STUDENT DATA

In recent years, there have been drastic improvements in the scope and the extent of data collection and management in schools. Student Information Systems (SIS), which originated as electronic grade management systems can now collect information related to student demographics, attendance, health, discipline, etc. The SISs can also store results of as many assessments as the schools choose to administer and enter. They have become data warehouses ripe for mining. The opportunity for the educational data mining community is that we have ready access to data that can help explore how student performance might be impacted by information already collected as noted above, e.g., student demographics and attendance, in highly complex and non-obvious ways.

Standardized tests are widely used by schools to identify the areas that they need to focus on, for their students to graduate. These tests are deliberately designed to be predictive of future student performance. However, there is no single test which is a true and representative indicator of the student performance. So, the schools administer several different tests from multiple testing agencies at different points of time over the school year. The test scores from these agencies are consolidated with the hope that some consistent patterns will emerge from the amalgamated data. Fig 2 shows the consolidated test scores for the students at a single grade level. All student and test names have been redacted for the purposes of privacy. The problem with the consolidated data is that it has high dimensionality and it can be noisy. Because of the high dimensionality, it is hard to find significant patterns in the correlation of the student performance that are consistent across multiple tests. Also, different tests might be measuring different attributes of the students e.g., some tests are strongly math oriented, while others could test both the math and language skills. Data mining techniques that operate on the latent attributes of the students that are derived from the test scores would be more accurate than the techniques that operate on the raw test scores.

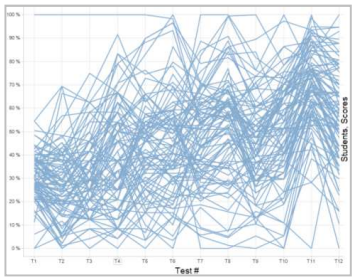

Fig 2: Consolidated test scores for students at a single grade level

A growing number of school districts are also adopting formative assessments to measure student skills in core subject areas throughout the school year. Fig 3 shows the results from formative assessment for the students in a single class. The main benefit of the formative assessments is that the teachers get more timely feedback about the effectiveness of their instructions and can adjust their instruction plans accordingly. To the data mining community, the availability of fine-grained data is both an opportunity as well as a challenge. By analyzing highly granular data, it is possible to detect actionable and otherwise non-obvious patterns of interest. However, the patterns might be hard to detect because of high dimensionality of the data.

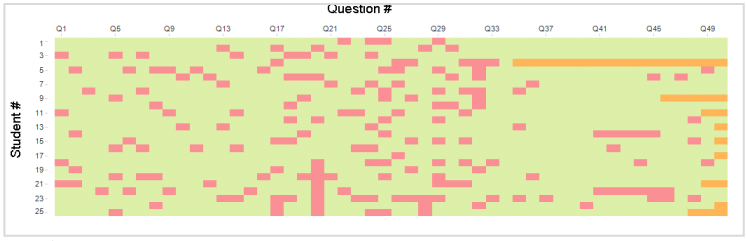

Fig 3: Results from a summative assessment: Green represents a correct response. Red and orange denote incorrect and skipped responses respectively.

2.1. Example Application – Co-clustering from Student Assessment Data

Clustering of students, based on their performance patterns, is a common application of student assessment data. In a typical classroom, different student groups have different learning needs and clustering can help identify these groups. However, the limitation of classical clustering algorithms (e.g. K-means, Hierarchical clustering) is that they can only identify student groups, but not the learning needs within the group. Co-clustering based approaches can be used to simultaneously discover student clusters as well as gaps in their learning (e.g. a set of questions that the students answered incorrectly or a specific concept which a set of students could not master). In order to identify the co-cluster from the data, we first represent the relationship between students and items as a bipartite graph, where edges are drawn between a student and an item only if the student performed a mistake on the item. The bipartite graph is then transformed into an adjacency matrix, which can be highly sparse. Spectral decomposition techniques are used to identify dense regions of the adjacency matrix. The "Spectrum" values of a graph are the Eigen vectors of a graph ordered on the strength of their corresponding Eigen values. The spectrum provides valuable insights into the connectivity of a graph. After the spectrum of a graph is identified, clustering algorithms can be applied to simultaneously identify a set of students and the items on which they had performed mistakes. It is also possible to apply hard clustering on students, to first divide them into disjoint sets, and then apply soft clustering on items, in order to assign the same items to multiple student clusters. An example of co-clusters discovered from student data is shown in Figure 4.

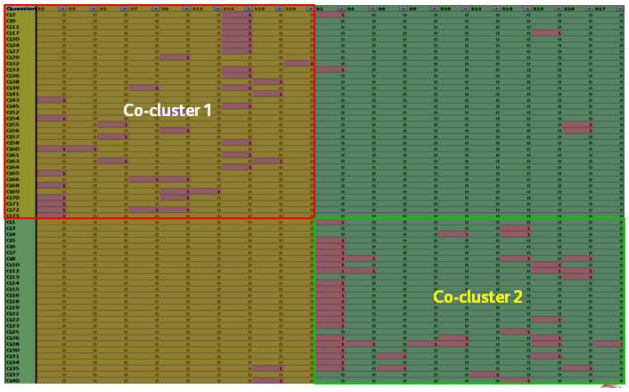

Fig 4: Two co-clusters discovered from student performance data on an assessment

3. KNOWLEDGE DISCOVERY FROM EDUCATIONAL DATA

In this section, we provide a few examples of actionable knowledge that can be extracted from the student achievement data. We also elaborate on the types of problems that the discovered knowledge would help address for the various agents involved in education.

3.1. Students: For most students, a better understanding of their strengths and weaknesses can help them stay motivated in the task of learning. A feedback on their problem areas and a set of specific intervention recommendations would help them take the necessary actions to redress the gaps in their learning. However, in the "one teacher, many learners" setting of today's classrooms, teachers find it extremely challenging to be cognizant of the learning needs of each individual student in the class. By mining the student performance data, it is possible to learn patterns of interest in the performance of the students. Automated reports that inform the students of their weaknesses, and at the same time apprise them of their strengths can engage students in taking responsibility of their own learning.

3.2. Teachers: Often teachers have to deal with vast amounts of information regarding student performance that is made available to them on a regular basis. Though the Student Information Systems have improved tremendously over the past few years, not everyone has access to the input or export of data in the system, however, and often a data manager works to do that off-site from the end users (teachers/admin). Because of the non-instant access to the data, many administrators and teachers have created do-it-yourself systems using basic spreadsheet software to store student data. In the absence of quality tools, they can make assumptions about the data that are not valid or reliable. One such problem for which they make use of their intuition is the task of clustering students. Identifying clusters [Jain98] based on the performance data of students gives the teachers an opportunity to encourage students with similar educational needs to work together and also to target instruction and materials meeting their specific needs. Subspace clustering algorithms [Cheng 2000] can be used to simultaneously discover the student clusters and the problem areas specific to these clusters.

Another problem that the teachers regularly face is that of finding an accurate diagnosis for the learning problems faced by the students. To remedy the issues caused by gaps in the student learning, longitudinal data of student performance can be mined. Bayesian inference based approaches [Ghahramani 98], in which the observations are used to calculate the probability that a hypothesis may be true, can used to accurately identify the root cause of a problem.

3.3. School Administrators: The school administrators would be most interested in analyzing the performance of their schools in comparison with other similar schools based on certain criteria. A common pitfall in analysis such as these is to just compare the average or the median values of the subjects in one group with that in another group. However, the group as a whole might have several component sub-populations, which makes summary statistics non-representative. Thus, first identifying clusters of students based on the independent variables, e.g., demographics, and then comparing similar clusters based on the dependent variables, e.g., student performance can show whether there is a true difference in the performance between the two clusters.

Many school administrators are also proactive in identifying students who would need special assistance in order to improve their performance. Algorithms that help in early detection of anomalous behavior [Zhang 08] from the performance data can help to identify such students.

3.4. District and State Level Administrators: The district and state level administrators often want to utilize the student achievement data for guided decision making and to enact systemic changes to the schools under their purview. They often have broad-ranging questions regarding issues such as gaps in the curriculum, correlation between student performance and other explanatory variables such as demographics, teaching methods, health and attendance records. Because of the manner in which data is combined and analyzed today, they often propose far-reaching changes based on limited evidence. Data analysis techniques that are robust to noisy and incomplete data and that can generate high quality actionable knowledge can be a value-add to these administrators. Ensemble techniques [Polikar 06] can help avoid overfitting and can combine multiple base models resulting in an accuracy that is higher than that generated by a single best model. Such techniques can provide unerring and reliable insights into the data for the administrators.

4. NEXT STEPS

Our work on educational data mining at Xerox Research is focused on data-driven methodologies to help individualize education. We are collaborating with teachers and school administrators to identify their biggest challenges and have also started a concerted effort of applying data mining algorithms on educational data. Xerox has a history of open innovation to solve challenging problems, and in this spirit, we are interested in collaborating with academic partners in order to find innovative solutions to the challenges in education.

REFERENCES

ALEVEN, V., MCLAREN, B., ROLL, I., KOEDINGER, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence and Education*, 16, 101-128.

BAKER, R.S.J.d., CORBETT, A.T., KOEDINGER, K.R. 2007. The Difficulty Factors Approach to the Design of Intelligent Tutoring Systems. International Journal of Artificial Intelligence in Education, 17 (4), 341-369.

BLACK, P., & WILLIAM, D. 1998. Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan, 80(2): 139-149

CHENG, Y., CHURCH, G.M . 2000. Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology: 93–103.

GHAHRAMANI, Z. 1998. Learning dynamic Bayesian Networks, Adaptive Processing of sequences and data structures

HORD, S.M. 1997. Professional Learning Communities: What are they and why are they important? Issues about Change. 6(1)

JAIN A.K., MURTY M.N., FLYNN P.J. 1999. Data clustering a review, ACM Computing Surveys, Volume 31, No 3

KAFTAN M.J., BUCK. A.G., HAACK. A. 2006. Using Formative Assessments to Individualize Instruction and Promote Learning

POLIKAR, R. (2006). "Ensemble based systems in decision making". IEEE Circuits and Systems Magazine

STEINBACH, M., ERTÖZ, l., KUMAR, V. 2003. The Challenges of Clustering High Dimensional Data, In New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition

Zhang. J., Gao. Q., Wang. H,. 2008. Anomaly detection in high-dimensional network data streams: A case study, Intelligence and Security Informatics