International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Data Analysis and Knowledge Extraction in Educational Data Mining (EDM)

K. Sirisha[1], Pathuri Siva Kumar[2]

[1, 2]Computer Science and Engineering, Rise Group of Institutions, Ongole, India

Abstract: ***While the field of educational data mining (EDM) has generated many innovations for improving educational software and student learning, the mining of student data has recently come under a great deal of scrutiny. Many stakeholder groups, including public officials, media outlets, and parents, have voiced concern over the privacy of student data and their efforts have garnered national attention. The momentum behind and scrutiny of student privacy has made it increasingly difficult for EDM applications to transition from academia to industry. Based on experience as academic researchers transitioning into industry, we present three primary areas of concern related to student privacy in practice: policy, corporate social responsibility, and public opinion. Our discussion will describe the key challenges faced within these categories, strategies for overcoming them, and ways in which the academic EDM community can support the adoption of innovative technologies in large-scale production.***

Keywords: Student privacy, student data, policy

1. Introduction

Educational data mining (EDM) is chiefly defined by the application of sophisticated data mining techniques to solving problems in education [1]. A powerful tool, EDM has been successfully incorporated into applications that optimize student learning in both research and commercial products. EDM's proven effectiveness has led many—from the U.S. government to individual teachers—to recognize the ability of student data in guiding education and to support the development and use of these technologies in schools. Consequently, applications utilizing EDM technologies have become more prevalent in school systems [2], [3].

However, the increase in EDM usage has raised public awareness of how much data is being collected about students. The applications and companies that collect and use student data are coming under scrutiny, as parents, advocates, and public officials grow concerned over student privacy. A recent cascade of events has focused attention on privacy concerns [4]. For example, there has been a rise in high-profile attacks on consumer data from online retailers and financial institutions. Large, well-trusted institutions have been targeted for using student data in undesirable ways [5]. Promising companies driven by student data have been brought down by public opinion with no evidence of wrong-doing. Calls for stricter policy from privacy advocates have led to more than 100 bills being introduced in U.S. state legislatures to address issues of student privacy in 2014 [4]. In response, the White House has announced plans for federal legislation modeled after state policies [6].

Negative media attention and increased legislation threaten to stifle EDM, particularly in commercial settings. Public opinion may make organizations wary to invest in and use EDM techniques while legislation could make it more difficult to collect and use student data in effective ways. We believe it is an incredibly important time for the EDM community to be aware of the challenges being faced in industry. The rise of concern over student privacy has strong implications for how new EDM approaches can be integrated into wide-reaching applications as well as the amount of funding available to public and private entities wishing to innovate in this space.

These issues are receiving rapidly increasing attention and driving action at the national level. It is critical that the discussions around these issues include experts from the EDM community. This paper discusses the issues and implications faced by commercial applications of educational data mining because of recent focus on student privacy. In this paper, we discuss the role of policy, corporate social responsibility, and public opinion in framing the work of and challenges to industry. We discuss strategies for overcoming these challenges and present opportunities for the EDM community to address rising concerns.

2. EDM and Industry

The profile of the EDM community has risen in the past decade— in research, commercial products, public attention—bolstered by three related shifts. First, educational technology has been more widely adopted. School systems are investing in laptops, mobile devices and other technologies in favor of static textbooks. These technologies offer opportunities for data collection that did not exist before. Student records are also increasingly digitized including test scores, attendance records, and bus schedules. These digitized records have generated a wealth of longitudinal data that was previously difficult and expensive to collect [7].

Second, there has been a dramatic rise in computational power and storage capacities. This storage allows for the collection and housing of large amounts of data, even data that is not presently known to be useful. The increased computational power has generated sophisticated algorithms that can mine large corpora of data to identify connections that would previously be impossible [8] and has even created the possibility for robust decision engines to operate in real time learning systems.

Finally, public officials and industry experts are starting to recognize the power of educational data mining [9]. Government funding opportunities for data-driven education solutions are on the rise, and reports estimate that educational data mining has the potential to provide meaningful economic impact worldwide [10].

There are many areas of EDM research, each with unique applications to industry. At the individual level, data on student behavior, from mouse clicks to eye tracking, provide insight on how students interact with educational technology. For example, EDM has produced models of help abuse [11], attention to hints [12], and conversational dynamics in online forums [13]. These insights and techniques can help commercial educational technology providers design better applications that support positive interactions with students while being user-friendly.

Another key area of research at the individual level is assessment. EDM applications have been used to identify student mastery as well as knowledge gaps. Frequently, these models are based on student performance on relevant tasks but can go beyond measuring what students did correctly and incorrectly by modeling underlying knowledge [14]. Some assessments are cleverly hidden, called ―stealth assessment,‖ in games or ot her non-threatening applications[15]. These systems develop robust models of student knowledge while avoiding the negative effects associated with test performance; in fact, students may not even knowing they are being tested. These techniques have import ant implications for educational technologies, ranging from the design of new systems that can revolutionize the way assessment is done in formal learning environments, to technologies that can identify gaps in student knowledge and recommend resources to help fill them.

EDM technologies have also driven personalized learning beyond tailoring instruction to what students know, but also to how they learn based on needs and preferences. Systems can identify commonly used strategies by students and select which are most effective, for particular individuals, under specific circumstances [16]. EDM techniques have also supported technologies that guide students towards learning how to regulate their own learning, by helping them to recognize and overcome weaknesses in their current approaches [17]. These techniques are critical in creating applications that use the most effective techniques and support personalized learning.

Finally, EDM research has examined mining data at higher levels, including schools and districts, for a variety of purposes such as exploring college readiness [18], identifying the best teachers[19], or driving district spending [7]. Commercial products are commonly used to house this level of data and communicate findings to necessary stakeholders. Data mining on this organizational or even regional level has allowed for the development of early warning systems to predict student drop-out before it happens as well as identify holes in district-level education [7].

In essence, ―educational data mining and learning analytics have the potential to make visible data that have heretofore gone unseen, unnoticed, and, therefore, unactionable‖ [9]. The approaches outlined in this section offer significant promise in helping to improve education delivery and outcomes, but their success is contingent on the collection, storage, and use of large amounts of quality student data. Companies who wish to collect and use student data must operate under increased public and governmental scrutiny, which can, and has, created barriers to the use of EDM in industry.

3. Role of the EDM Community

The barriers to industry applications of educational data mining techniques stem from several sources. Existing and proposed policy put restrictions on how data can be collected, stored and used. Companies can technically comply with legislation without much impact on their product or processes. However, strictly adhering to policies and offering real privacy protection often makes accessing and using educational tools more difficult, giving less socially responsible companies a competitive advantage. Public opinion can lead to the destruction of companies with no unethical practices and can drive money away from investment in data-based educational technologies. The EDM community has an important role to play in keeping these challenges in check and allowing innovation to thrive (Table 1).

3.1 Transparency

A lack of clarity, rampant misunderstanding, and a high degree of uncertainty fuel sentiment against the collection and use of student data. The main concerns of many parents and privacy advocates are largely not reflective of actual practice.

Consequently, the EDM community is unique positioned to advance public understanding for what student data is really being used. EDM professionals can better describe how data is being used, what innovations it supports, explain the focus of current research, and portray likely research foci of the field. Parental concerns may be allayed knowing that people are not actively contributing to the outcomes they most fear.

The community can also disseminate details about the effectiveness of these approaches beyond the research community. Showing the strengths of these techniques may help concerned individuals see the benefits that individual children and the education system as a whole stand to gain.

As new approaches are developed, consider creating public-facing talking points that can be used to communicate with concerned parties. These points should describe what data is being used and how it can benefit students. They should be written in a way that is clear and easy for non-experts to understand.

Table 1: The role of the EDM community on the issue of student privacy

Point of Concern	Proposed Solution	Action Item
Policy	Policy Activism	•Remain abreast of proposed or approved policy Changes • Actively voice expert opinionsto policy makers.
Corporate Social Responsibility	Awareness of classroom	• constraints Develop algorithms that minimize the amount of data needed to produce effective results where possible. • Avoid requirements for individual accounts when possible
Understanding public opinion	Public Opinion Transparency	• Actively work to correct misconceptions about student data and pnvacy concems • Set research agendas aimed at better understanding publicunderstanding of privacy issues.

3.2 Research Agendas

The EDM community can also drive research towards areas that may help compliance with legislation and improve public opinion. Algorithms that minimize the amount of data needed to produce effective results would be beneficial to companies wishing to keep privacy concerns at bay. Researchers should consider the tradeoffs when developing new ―big data" approaches. More data may lead to more effective techniques but it also may represent an increased violation of privacy. Finding a balance can support widespread dissemination in commercial technologies.

It is important that researchers understand the classroom constraints of commercial educational technologies, especially when it comes to privacy. For a variety of reasons it is often less feasible to guarantee that data comes from a specific individual. Approaches that are robust enough to take this into account will allow educational technologies to be successful in more environments.

An additional area of research that could benefit from the involvement of the EDM community is research on the public understanding of privacy issues. The EDM community could be involved in cross-disciplinary research to ensure that communication surrounding EDM techniques is accurate and clear, and organizational privacy policies are widely understood.

3.3 Policy Activism

Finally, we encourage members of the EDM community to become active as policy debates grow. It is important to stay up to date on proposed policy changes and to consider how these changes may impact research agendas and the commercial applicability of those findings. Policy changes may increase constraints in commercial applications that could drive shifts in funding made available to EDM research. The policy changes affect both communities.

The discussion also needs more contributions from EDM experts. Consider voicing concerns to local officials and provide guidance on how policy should be directed. Too much of the current dialogue is based on a fear and misunderstanding. These voices are currently overpowering the experts who support the use of data in education.

4. Conclusion

Educational data mining offers significant promise in improving student learning and education systems as a whole. However, these systems are often driven by the collection of large amounts of student data, which is a growing concern to many. Shifts in public opinion and policy have led to barriers to the adoption of EDM technologies in commercial applications and threaten to stifle future innovation. Several fundamental issues are driving this trend.

The first is the role of trust, fear, and misunderstanding. It is difficult to combat the fear associated with the unknown. Companies and experts in the field must work hard to both gain the trust of the public and communicate what is actually being done with student data. Trust must extend the other way as well. Companies need to trust that by being open about their practices they will not be attacked by concerned external stakeholders. Fear from companies about the reactions of privacy advocates encourages silence on their parts and serves to reduce overall transparency. Both parties must build trust to move towards an open and productive dialogue.

Another recurring theme centers on legislation that has not yet had the desired effect. Privacy advocates view current legislation as too lenient and many companies are able to comply without actually protecting student data. In fact, the legislation may actually harm companies that do the most to protect student privacy. Voluntary pledges offer one solution, though they are not without problems; conflicts of interest often erode even the best self-policing strategies. Many, if not most, companies may support the spirit of such pledges but be unable to sign due to any number various technicalities. Active involvement from all invested parties will be crucial to designing new legislation that will strike a balance between allowing data to be used for the good of education, while protecting the privacy of individual students.

Finally, differing views on the appropriateness of private institutions delivering public goods underscore many of the issues discussed. If commercial vendors are going to be the major providers of educational technologies to school systems there needs to be a shift in how the public perceives these companies. Stifling the success of these companies only serves to keep innovative learning technologies out of the classroom. Still, deference to privacy concerns is an important component of occupying a role in part characterized by public stewardship. Discussions about the ethical limits of financially profiting off of student data need to be addressed directly by corporate, research, and public interests with adequate emphasis on risk and potential system improvements

Overall, there are variety of issues contributing to concerns over student privacy and how these concerns impact industry applications of educational data mining. These issues are extremely prominent and are not expected to lose momentum soon. The EDM community stands to play an important role in how discussions and legislation around student privacy evolve in the coming years. The landscape

of educational data and privacy will continue to shift, and we hope with increased involvement this shift will be positive for researchers and industries interested in using educational data mining to support student learning.

References

[1] G. Siemens and R. S. J. Baker, —Learning Analytics and Educational Data Mining : Towards Communication and Collaboration," pp. 252–254, 2012.
[2] S. Simon, —Data Mining Your Children," Politico, 15-May- 2014.
[3] N. Singer, —With Tech Taking Over in Schools, Worries Rise," The New York Times, 14-Sep-2014.
[4] S. Trainor, —Student data privacy is cloudy today, clearer tomorrow," Phi Delta Kappan, vol. 96, no. 5, pp. 13–18, 2015.
[5] B. Herold, —inBloom to Shut Down Amid Growing DataPrivacy Concerns," Education Week, 04-Feb-2014.
[6] —FACT SHEET: Safeguarding American Consumers & Families," The White House, 2015. [Online]. Available: http://www.whitehouse.gov/the-press-office/2015/01/12/factsheet-safeguarding-american-consumers-families.
[7] J. McQuiggan and A. W. Sapp, Implement, Improve and Expand Your Statewide Longitudinal Data System: Creating a Culture of Data in Education. 2014.
[8] Mayer-Schonberger and K. Cukier, Big Data. New York, New York: Haughton Mifflin Harcourt Publishing Company, 2013.
[9] U. S. D. of Education, —Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief," 2012.
[10] J. Manyika, M. Chui, D. Farrel, S. Van Kuiken, P. Groves, and E. Almasi, —Open data: Unlocking Innovation and Performance with Liquid Information," 2013.
[11] V. Aleven and K. Koedinger, —Limitations of Student Control: Do Students Know When They Need Help?," in Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 2000, pp. 292–303. [12] C. Conati, N. Jaques, and M. Muir, —Understanding attention to adaptive hints in educational games: an eye-tracking study," Int. J. Artif. Intell. Educ., vol. 23, pp. 136–161, 2013.
[12] M. Wen, D. Yang, and C. Rose, —Sentiment Analysis in MOOC Discussion Forums: What does it tell us?," in Proceedings of the 7th International Conference on Educational Data Mining, 2014, pp. 257–260.
[13] R. S. J. Baker, A. T. Corbett, and V. Aleven, —More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing," Knowl. Creat. Diffus. Util., pp. 406–415, 2008.
[14] V. Shute, —Stealth Assessment in Computer-Based Games to Support Learning," in Computer Games and Instruction, 2011, pp. 503–523.
[15] J. P. Rowe, L. R. Shores, B. W. Mott, and J. C. Lester, —Integrating Learning, Problem Solving, and Engagement in Narrative-Centered Learning Environments," Int. J. Artifical Intell. Educ., vol. 21, no. 1–2, pp. 115–133, 2011.
[16] J. Sabourin, L. R. Shores, B. W. Mott, and J. C. Lester, —Predicting Student Self-Regulation Strategies in GameBased Learning Environments," in Proceedings of the 11th International Conference on Intelligent Tutoring Systems, 2012.
[17] H. Chen, —Identifying Early Indicators for College Readiness," 2007.
[18] L. Pappano, —Using Research to Predict Great Teachers," Harvard Education Letter, 2011.
[19] M. Sieber, J. Tolich. —Planning ethically responsible research" Sage Publications, 2012.
[20] S. Carey, —Students, Parents and the School Record Prison A Legal Strategy for Preventing Abuse.pdf," J. Law Educ., vol. 3, p. 365, 1974.
[21] T. L. Elliott, D. Fatemi, and S. Wasan, —Student Privacy Rights — History , Owasso , and FERPA," J. High. Educ. Theory Pract., vol. 14, no. 4, 2014.
[22] J. R. Reidenberg, N. C. Russell, J. Kovnot, T. B. Norton, and R. Cloutier, —Privacy and Cloud Computing in Public Schools," 2013.
[23] R. Silverblatt, —Hiding behind ivory towers: Penalizing schools that improperly invoke student privacy," Georgetown Law J., vol. 101, pp. 493–517, 2013.
[24] B. Smith and J. Mader, —Protecting Students' Privacy - By Law," Sci. Teach., vol. 81, no. December, 2014.
[25] Children's Online Privacy Protection Act of 1998, 5 U.S.C. 6501-6505.
[26] A. Allen, —Minor Distractions: Children, Privacy and Ecommerce," Houston Law Review, 2001.
[27] J. Mayfield, —Revised Children's Online Privacy Protection Rule Goes Into Effect Today Federal Trade Commission," Federal Trade Commission, 01-Jul-2013.
[28] Data Quality Campaign, —2014 Student Data Privacy Bills," 2014.
[29] E. Brown, —Obama to propose new student privacy legislation," The Washington Post, Washington D.C., 19-Jan- 2015.
[30] S. Simon, —Barack Obama to seek limits on student data mining," Politico, 11-Jan-2015.
[31] H. Tsukayama, —More than 70 companies just signed a pledge to protect student data priacy - with some notable exceptions," The Washington Post, 12-Jan-2015.
[32] D. Banisar, —Privacy and data protection around the wolrd," in 21st International Conference on Privacy and Personal Data Protection, 1999.
[33] G. Yee, —Security and Privacy in Distance Education," in Information Security and Ethics: Concepts, Methodologies, Tools, and Applications, 1st ed., H. Namati, Ed. 2007, p. 4110.
[34] D. Boyd, —Response to COPPA Rule Review, 16 CFR part 312, Project No. P-104503," Washington D.C., 2011.
[35] N. S. B. Association, —Data in the Cloud: A Legal and Policy Guide for School Boards on Student Data Privacy in the Cloud Computing Era," Alexandria, VA, 2014.
[36] C. S. Media, —Student Privacy Survey," 2014.
[37] S. Fox, —In the Matter of COPPA Rule Review, 16 CFR Part 312, Project No. P-104503," Washington D.C., 2011.

[38] J. Palfrey, D. Boyd, and U. Gasser, ―How the COPPA, as Implemented, Is Misinterpreted by the Public: A Research Perspective," 2010.
[39] Pew Research Center, ―What Internet Users Know about Technology and the Web," 2014.
[40] C. DeLorme, ―Response to COPPA Rule: Comments to be placed on the public record," Washington D.C., 2012.
[41] M. Madden, S. Cortesi, U. Gasser, A. Lenhart, and M. Duggan, ―Parents, Teens, and Online Privacy," 2012.
[42] B. L. Wilcox, D. Kunkel, J. Cantor, P. Dowrick, S. Linn, and E. Palmer, ―Report of the APA Task Force on Advertising and Children," 2004.
[43] K. Vanlehn, ―The Behavior of Tutoring Systems," Int. J. Artif. Intell. Educ., vol. 16, no. 3, pp. 227–265, 2006.