Half Title Page

Series Page

Title Page

LOC Page

To my wife Ana and my son Cristóbal

Cristóbal Romero

To my wife Inma and my daugther Marta

Sebastián Ventura

To my wife Ekaterina and my daugther Aleksandra

Mykola Pechenizkiy

To my wife Adriana and my daugther Maria

Ryan S. J. d. Baker

Contents

Part II Case Studies

Preface

The Purpose of This Book

The goal of this book is to provide an overview of the current state of knowledge of educational data mining (EDM). The primary goal of EDM is to use large-scale educational data sets to better understand learning and to inform the learning process. Although researchers have been studying human learning for over a century, what is different about EDM is that it makes use not of experimental subjects learning a contrived task for 20 min in a lab setting; rather, it typically uses data from students learning school subjects, often over the course of an entire school year. For example, it is possible to observe students learning a skill over an eight-month interval and make discoveries about what types of activities result in better long-term learning, to learn about the impact of what time students start their homework has on classroom performance, or to understand how the length of time students spend reading feedback on their work impacts the quality of their later efforts.

In order to conduct EDM, researchers use a variety of sources of data such as intelligent computer tutors, classic computer-based educational systems, online class discussion forums, electronic teacher gradebooks, school-level data on student enrolment, and standardized tests. Many of these sources have existed for decades or, in the case of standardized testing, about 2000 years. What has recently changed is the rapid improvement in storage and communication provided by computers, which greatly simplifies the task of collecting and collating large data sets. This explosion of data has revolutionized the way we study the learning process.

In many ways, this change parallels that of bioinformatics 20 years earlier: an explosion of available data revolutionized how much research in biology was conducted. However, the larger number of data was only part of the story. It was also necessary to discover, adapt, or invent computational techniques for analyzing and understanding this new, vast quantity of data. Bioinformatics did this by applying computer science techniques such as data mining and pattern recognition to the data, and the result has revolutionized research in biology. Similarly, EDM has the necessary sources of data. More and more schools are using educational software that is capable of recording for later analysis every action by the student and the computer. Within the United States, an emphasis on educational accountability and high stakes standardized tests has resulted in large electronic databases of student performance. In addition to these data, what we need are the appropriate computational and statistical frameworks, techniques to make sense of the data, as well as researchers to ask the right questions of the data.

No one discipline has the necessary expertise to conduct EDM research. Thus, the community, as can be seen by the chapter authors of this book, is composed of people from multiple disciplines. Computer science provides expertise in working with large quantities of data, both in terms of machine learning and data-mining techniques that scale gracefully to data sets with millions of records, as well as address real-world concerns such as "scrubbing" data to ensure systematic errors in the source data do not lead to erroneous results. Statisticians and psychometricians provide expertise in understanding how to properly analyze complex study designs, and properly adjust for the fact that most educational data are not from a classic randomized controlled study. These two communities are strong in statistical and computational techniques, but techniques and data are not sufficient to advance a scientific domain; researchers with basic understanding of the teaching and learning process are also required. Thus, education researchers and psychologists are key participants in the EDM community.

Main Avenues of Research in Educational Data Mining

There are three major avenues of research in EDM. They nicely align with the classic who–what–where–when interrogatives.

The first avenue is work on developing computational tools and techniques, determining which ones are best suited to working with large educational data sets, and finding best practices for evaluation metrics and model fitting. Examples of such efforts include experimenting with different visualization techniques for how to look at and make sense of the data. Since educational data sets are often longitudinal, encompassing months and sometimes years, and rich interactions with the student can occur during that time, some means of making sense of the data is needed. Another common approach in EDM is using variants of learning curves to track changes in student knowledge. Learning curves are some of the oldest techniques in cognitive psychology, so EDM efforts focus on examining more flexible functional forms, and discovering what other factors, such as student engagement with the learning process, are important to include. One difficulty with complex modeling in EDM is there is often no way of determining the best parameters for a particular model. Well-known techniques such as hill climbing can become trapped in local maxima. Thus, empirical work about which model-fitting techniques perform well for EDM tasks is necessary.

This work on extending and better understanding our computational toolkit is a necessary foundation to EDM. Work in this area focuses on how we can extract information from data. At present, although a majority of EDM research is in this avenue, the other two are not less important—just less explored.

The second avenue is determining what questions we should ask the data. There are several obvious candidates: Does the class understand the material well enough to go on? Do any students require remedial instruction? Which students are likely to need academic counseling to complete school successfully? These are questions that have been asked and answered by teachers for millennia. EDM certainly enables us to be data driven and to answer such questions more accurately; however, EDM's potential is much greater. The enormous data and computational resources are a tremendous opportunity, and one of the hardest tasks is capitalizing on it: what are new and interesting questions we can answer by using EDM? For example, in educational settings there are many advantages of group projects. Drawbacks are that it can be hard to attribute credit and, perhaps more importantly, to determine which groups are having difficulties—perhaps even before the group itself realizes. A tool that is able to analyze student conversations and activity, and automatically highlight potential problems for the instructor would be powerful, and has no good analog in the days before computers and records of past student collaborations were easily available. Looking into the future, it would be useful if we could determine if a particular student would be better served by having a different classroom teacher, not because one teacher is overall a better choice, but because for this type of student the teacher is a better choice. The first example is at the edge of what EDM is capable; the second is, for now, beyond our capabilities.

This job of stretching our horizons and determining what are new, exciting questions to ask the data is necessary for EDM to grow.

The third avenue of EDM is finding who are educational stakeholders who could benefit from the richer reporting made possible with EDM. Obvious interested parties are students and teachers. However, what of the student's parents? Would it make sense for them to receive reports? Aside from report cards and parent–teacher conferences, there is little communication to parents about their child's performance. Most parents are too busy for a detailed report of their child's school day, but what of some distilled information? A system that informed parents if their child did not complete the homework that was due that day could be beneficial. Similarly, if a student's performance noticeably declines, such a change would be detectable using EDM and the parents could be informed. Other stakeholders include school principals, who could be informed of teachers who were struggling relative to peers, and areas in which the school was performing poorly. Finally, there are the students themselves. Although students currently receive an array of grades on homeworks, quizzes, and exams, they receive much less larger-grain information, such as using the student's past performance to suggest classes to take, or that the student's homework scores are lower than expected based on exam performance. Note that such features also change the context of educational data as something that is used in the classroom, to something that is potentially used in a completely different place.

Research in this area focuses on expanding the list of stakeholders for whom we can provide information, and where this information is received. Although there is much potential work in this area that is not technically demanding, notifying parents of missed homework assignments is simple enough, such work has to integrate with a school's IT infrastructure, and changes the ground rules. Previously teachers and students controlled information flow with parents, now parents are getting information directly. Overcoming such issues is challenging. Therefore, this area has seen some attention, but is relatively unexplored by EDM researchers.

The field of EDM has grown substantially in the past five years, with the first workshop referred to as "Educational data mining" occurring in 2005. Since then, it has held its third international conference in 2010, had one book published, has its own online journal, and is now having this book published. This growth is exciting for multiple reasons. First, education is a fundamentally important topic, rivaled only by medicine and health, which cuts across countries and cultures. Being able to better answer age-old questions in education, as well as find ways to answer questions that have not yet been asked, is an activity that will have broad impact on humanity. Second, doing effective educational research is no longer about having a large team of graduate assistants to score and code data, and having sufficient offices with filing cabinets to store the results. There are public repositories of educational data sets for others to try their hand at EDM, and anyone with a computer and Internet connection can join the community. Thus, a much larger and broader population can participate in helping improve the state of education.

This book is a good first step for anyone wishing to join the EDM community, or for active researchers wishing to keep abreast of the field. The chapters are written by key EDM researchers, and cover many of the field's essential topics. Thus, the reader gets a broad treatment of the field by those on the front lines.

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01[illegible]2098 USA
Tel: 508 647 7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Joseph E. Beck
Worcester Polytechnic Institute, Massachusetts

Editors

Dr. Cristóbal Romero is an associate professor in the Department of Computer Science at the University of Córdoba, Spain. His research interests include applying artificial intelligence and data-mining techniques in education and e-learning systems. He received his PhD in computer science from the University of Granada, Spain, in 2003. The title of his PhD thesis was "Applying data mining techniques for improving adaptive hypermedia web-based courses." He has published several papers about educational data mining in international journals and conferences, and has served as a reviewer for journals and as a PC member for conferences. He is a mem- AQ1
ber of the International Working Group on Educational Data Mining and an organizer or PC member of conferences and workshops about EDM. He was conference chair (with Sebastián Ventura) of the Second International Conference on Educational Data Mining (EDM'09).

Dr. Sebastián Ventura is an associate professor in the Department of Computer Science at the University of Córdoba, Spain. He received his PhD in sciences from the University of Córdoba in 2003. His research interests include machine learning, data mining, and their applications, and, recently, in the application of KDD techniques in e-learning. He has published several papers about educational data mining (EDM) in international journals and conferences. He has served as a reviewer for several journals such as *User Modelling and User Adapted Interaction*, *Information Sciences*, and *Soft Computing*. He has also served as a PC member in several research EDM forums, including as conference chair (with Cristóbal Romero) of the Second International Conference on Educational Data Mining (EDM'09).

Dr. Mykola Pechenizkiy is an assistant professor in the Department of Computer Science, Eindhoven University of Technology, the Netherlands. He received his PhD in computer science and information systems from the University of Jyväskylä, Finland, in 2005. His research interests include knowledge discovery, data mining, and machine learning, and their applications. One of the particular areas of focus is on applying machine learning for modeling, changing user interests and characteristics in adaptive hypermedia applications including, but not limited to, e-learning and e-health. He has published several papers in these areas, and has been involved in the organization of the conferences, workshops, and special tracks.

K10250_C000.indb 15 5/28/2010 8:58:23 AM

Dr. Ryan S. J. d. Baker is an assistant professor of psychology and the learning sciences in the Department of Social Science and Policy Studies at Worcester Polytechnic Institute, Massachusetts, with a collaborative appointment in computer science. He graduated from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2005, with a PhD in human–computer interaction. He was a program chair (with Joseph Beck) of the First International Conference on Educational Data Mining, and is an associate editor of the *Journal of Educational Data Mining* and a founder of the International Working Group on Educational Data Mining. His research is at the intersection of educational data mining, machine learning, human–computer interaction, and educational psychology, and he has received five best paper awards or nominations in these areas. He is the former technical director of the Pittsburgh Science of Learning DataShop, the world's largest public repository for data on the interaction between students and educational software.

Contributors

Wil van der Aalst
Department of Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, the Netherlands

Analía Amandi
AQ2 ISISTAN Research Institute
UNICEN University
Tandil, Argentina

AQ3 **Saleema Amershi**
University of Washington
Seattle, Washington

Brigham S. Anderson
Worcester Polytechnic Institute
Worcester, Massachusetts

and

Carnegie Mellon University
Pittsburgh, Pennsylvania

Cláudia Antunes
Department of Information Systems and Computer Science
Instituto Superior Técnico
Lisbon, Portugal

Ivon Arroyo
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts

Ryan S. J. d. Baker
Department of Social Science and Policy Studies
Worcester Polytechnic Institute
Worcester, Massachusetts

Tiffany Barnes
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, North Carolina

Joseph E. Beck
Computer Science Department
Worcester Polytechnic Institute
Worcester, Massachusetts

and

Carnegie Mellon University
Pittsburgh, Pennsylvania

Javier Bravo
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Winslow Burleson
Department of Computer Science and Engineering
Arizona State University
Tempe, Arizona

Toon Calders
Eindhoven University of Technology
Eindhoven, the Netherlands

Cristina Carmona
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga
Málaga, Spain

Gladys Castillo
Department of Mathematics
Centre for Research on Optimization and Control
University of Aveiro
Aveiro, Portugal

Carlos de Castro
Department of Computer Science and Numerical Analysis
University of Cordoba
Cordoba, Spain

Félix Castro
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Spain

and

Centro de Investigación en Tecnologías de Información y Sistemas
Universidad Autónoma del Estado de Hidalgo
Hidalgo, Mexico

Mihaela Cocea
London Knowledge Lab
The University of London
London, United Kingdom

and

School of Computing
National College of Ireland
Dublin, Ireland

Cristina Conati
University of British Columbia
Vancouver, British Columbia, Canada

David G. Cooper
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts

Marvin Croy
Department of Philosophy
University of North Carolina at Charlotte
Charlotte, North Carolina

Andrew Cuneo
Carnegie Mellon University
Pittsburgh, Pennsylvania

Kyle Cunningham
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Sidney D'Mello
Institute for Intelligent Systems
The University of Memphis
Memphis, Tennessee

Philippe Fournier-Viger
Department of Computer Science
University of Quebec in Montreal
Montreal, Quebec, Canada

Enrique García
Department of Computer Science and Numerical Analysis
University of Cordoba
Cordoba, Spain

Daniela Godoy
ISISTAN Research Institute
UNICEN University
Tandil, Argentina

Evandro Gouvea
Carnegie Mellon University
Pittsburgh, Pennsylvania

Art Graesser
Institute for Intelligent Systems
The University of Memphis
Memphis, Tennessee

Wilhelmiina Hämäläinen
University of Helsinki
Helsinki, Finland

Cristina L. Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts

and

Carnegie Mellon University
Pittsburgh, Pennsylvania

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts

and

Carnegie Mellon University
Pittsburgh, Pennsylvania

Cecily Heiner
Carnegie Mellon University
Pittsburgh, Pennsylvania

Arnon Hershkovitz
Knowledge Technology Lab
School of Education
Tel Aviv University
Tel Aviv, Israel

Earl Hunt
Department of Psychology
University of Washington
Seattle, Washington

Octavio Juarez
Carnegie Mellon University
Pittsburgh, Pennsylvania

Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania

Judy Kay
School of Information Technologies
University of Sydney
Sydney, New South Wales, Australia

Jihie Kim
Information Sciences Institute
University of Southern California
Marina del Rey, California

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Irena Koprinska
School of Information Technologies
University of Sydney
Sydney, New South Wales, Australia

Brett Leber
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Tara M. Madhyastha
Department of Psychology
University of Washington
Seattle, Washington

David Masip
Universitat Oberta de Catalunya
Barcelona, Spain

Manolis Mavrikis
London Knowledge Lab
The University of London
London, United Kingdom

Riccardo Mazza
Faculty of Communication Sciences
University of Lugano
Lugano, Switzerland

Gordon G. McCalla
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

Victor H. Menendez
Facultad de Matemáticas
Universidad Autónoma de Yucatán
Merida, Mexico

Engelbert Mephu Nguifo
Department of Computer Sciences
Université Blaise-Pascal Clermont 2
Clermont-Ferrand, France

Agathe Merceron
Media and Computer Science Department
Beuth University of Applied Sciences Berlin
Berlin, Germany

Eva Millán
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga
Malaga, Spain

Julià Minguillón
Universitat Oberta de Catalunya
Barcelona, Spain

Enric Mor
Universitat Oberta de Catalunya
Barcelona, Spain

Jack Mostow
Carnegie Mellon University
Pittsburgh, Pennsylvania

Rafi Nachmias
Knowledge Technology Lab
School of Education
Tel Aviv University
Tel Aviv, Israel

Àngela Nebot
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Spain

John C. Nesbit
Faculty of Education
Simon Fraser University
Burnaby, British Columbia, Canada

Roger Nkambou
Department of Computer Science
University of Quebec in Montreal
Montreal, Quebec, Canada

Alvaro Ortigosa
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Zachary A. Pardos
Worcester Polytechnic Institute
Worcester, Massachusetts

and

Carnegie Mellon University
Pittsburgh, Pennsylvania

Mykola Pechenizkiy
Department of Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, the Netherlands

Kaska Porayska-Pomsta
London Knowledge Lab
The University of London
London, United Kingdom

Manuel E. Prieto
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ciudad Real, Spain

Sujith Ravi
Information Sciences Institute
University of Southern California
Marina del Rey, California

Cristóbal Romero
Department of Computer Science and Numerical Analysis
Cordoba University
Cordoba, Spain

Richard Scheines
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Erin Shaw
Information Sciences Institute
University of Southern California
Marina del Rey, California

Judy Sheard
School of Computer Science and Software Engineering
Monash University
Melbourne, Victoria, Australia

Benjamin Shih
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Alida Skogsholm
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

John Stamper
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Tiffany Y. Tang
Konkuk University
Chungju-si, South Korea

Nikola Trčka
Department of Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, the Netherlands

Alfredo Vellido
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Spain

Sebastián Ventura
Department of Computer Science and Numerical Analysis
Cordoba University
Cordoba, Spain

César Vialardi
Facultad de Ingeniería de Sistemas
Universidad de Lima
Lima, Peru

Mikko Vinni
University of Eastern Finland
Joensuu, Finland

Stephan Weibelzahl
School of Computing
National College of Ireland
Dublin, Ireland

Philip H. Winne
Faculty of Education
Simon Fraser University
Burnaby, British Columbia, Canada

Beverly P. Woolf
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts

Yabo Xu
School of Computing Science
Simon Fraser University
Burnaby, British Columbia, Canada

Kalina Yacef
School of Information Technologies
University of Sydney
Sydney, New South Wales, Australia

Amelia Zafra
Department of Computer Science and Numerical Analysis
University of Cordoba
Cordoba, Spain

Alfredo Zapata
Facultad de Educación
Universidad Autónoma de Yucatán
Merida, Mexico

Mingming Zhou
Faculty of Education
Simon Fraser University
Burnaby, British Columbia, Canada

AUTHOR QUERIES

[AQ1] Please provide the expanded form of "PC" if appropriate.

[AQ2] Please expand "ISISTAN" and "UNICEN," if appropriate.

[AQ3] Please provide department name for Saleema Amershi, Brigham S. Anderson, Joseph E. Beck, Toon Calders, Cristina Conati, Andrew Cuneo, Kyle Cunningham, Sidney D'Mello, Daniela Godoy, Evandro Gouvea, Art Graesser, Wilhelmiina Hämäläinen, Cristina L. Heffernan, Neil T. Heffernan, Cecily Heiner, Octavio Juarez, Judy Kay, Jihie Kim, Kenneth R. Koedinger, Irena Koprinska, Brett Leber, David Masip, Gordon G. McCalla, Julià Minguillón, Enric Mor, Jack Mostow, Zachary A. Pardos, Sujith Ravi, Richard Scheines, Erin Shaw, Judy Sheard, Benjamin Shih, Alida Skogsholm, John Stamper, Tiffany Y. Tang, Mikko Vinni, Stephan Weibelzahl, Yabo Xu, and Kalina Yacef.