
IGI PUBLISHING **ITB15230**
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter XIII

Data Mining in the Social Sciences and Iterative Attribute Elimination

Anthony Scime, SUNY Brockport, USA
Gregg R. Murray, SUNY Brockport, USA
Wan Huang, SUNY Brockport, USA
Carol Brownstein-Evans, SUNY Brockport, USA

Abstract

Immense public resources are expended to collect large stores of social data, but often these data are under-examined thereby missing potential opportunities to shed light on some of society's pressing problems. This chapter proposes and demonstrates data mining in general and an iterative attribute-elimination process in particular as important analytical tools to exploit more fully these important data from the social sciences. We use an iterative domain-expert and data mining process to identify attributes that are useful for addressing distinct and nontrivial research issues in social science—presidential vote choice and living arrangement outcomes for maltreated children—using the American National Election Studies (ANES) from political science and the National Survey on Child and Adolescent

Well-Being (NSCAW) from social work. We conclude that data mining is useful for more fully exploiting important but under-evaluated data collections for the purpose of addressing some important questions in the social sciences.

Data Mining: Ethical, Theoretical, and Practical Motivations

[D]ata are expensive in terms of time, effort, money, and other resources… If the research was worth doing, the data are worth a thorough analysis, being held up to the light in many different ways so that our research participants, our funding agencies, our science, and society will all get their time and their money's worth. (Rosenthal, 1994, p. 130)

More than $2.4 billion was spent on social science research and development at colleges and universities alone in 2003 (Jankowski, 2005). This investment represents not only a financial expenditure, but also the expenditure of countless hours, days, weeks, and months of researcher, participant, and administrator time and effort and other exertions. The results of this immense investment are often embodied in extensive data sets such as the general social survey, which has collected data on American society and attitudes approximately every two years since 1975 (NORC, 2006), the Uniform Crime Reporting Program, which has gathered information on crime levels and law enforcement administration, operation, and management in the U.S. since 1930 (Federal Bureau of Investigation, 2004), and the Baccalaureate and Beyond Longitudinal Study, which has followed about 11,000 students who completed their baccalaureate degree in 1992-93 to assess their education and work experience (Wine, Cominole, Wheeless, Dudley, & Franklin, 2005).

Minimally, the desire for basic efficiency demands reasonable output from this substantial input into research and development. But the challenge accepted by social investigators is greater than simple efficiency. For example, the mission statement of the National Science Foundation indicates its objective is "to promote the progress of science; to advance the national health, prosperity, and welfare; [and] to secure the national defense…" (National Science Foundation, 2005). The challenge, then, is to resolve important issues, often calling for high risk, high payoff research. In these endeavors, large quantities of data often are collected, some of which may contain the keys to resolving important social issues but which remain inaccessible due to the inability to evaluate comprehensively and efficiently the mass of data. Given the resources invested and potential for important findings, then, there is an ethical imperative to exploit these data to their fullest extent.