Ethics of Data Mining

Jack Cook
Rochester Institute of Technology, USA

INTRODUCTION

Decision makers thirst for answers to questions. As more data is gathered, more questions are posed: Which customers are most likely to respond positively to a marketing campaign, product price change or new product offering? How will the competition react? Which loan applicants are most likely or least likely to default? The ability to raise questions, even those that currently cannot be answered, is a characteristic of a good decision maker. Decision makers no longer have the luxury of making decisions based on gut feeling or intuition. Decisions must be supported by data; otherwise decision makers can expect to be questioned by stockholders, reporters, or attorneys in a court of law. Data mining can support and often direct decision makers in ways that are often counterintuitive. Although data mining can provide considerable insight, there is an "inherent risk that what might be inferred may be private or ethically sensitive" (Fule & Roddick, 2004, p. 159).

Extensively used in telecommunications, financial services, insurance, customer relationship management (CRM), retail, and utilities, data mining more recently has been used by educators, government officials, intelligence agencies, and law enforcement. It helps alleviate data overload by extracting value from volume. However, data analysis is not data mining. Query-driven data analysis, perhaps guided by an idea or hypothesis, that tries to deduce a pattern, verify a hypothesis, or generalize information in order to predict future behavior is not data mining (Edelstein, 2003). It may be a first step, but it is not data mining. Data mining is the process of discovering and interpreting meaningful, previously hidden patterns in the data. It is not a set of descriptive statistics. Description is not prediction. Furthermore, the focus of data mining is on the process, not a particular technique, used to make reasonably accurate predictions. It is iterative in nature and generically can be decomposed into the following steps: (1) data acquisition through translating, cleansing, and transforming data from numerous sources, (2) goal setting or hypotheses construction, (3) data mining, and (4) validating or interpreting results.

The process of generating rules through a mining operation becomes an ethical issue, when the results are used in decision-making processes that affect people or when mining customer data unwittingly compromises the privacy of those customers (Fule & Roddick, 2004). Data miners and decision makers must contemplate ethical issues before encountering one. Otherwise, they risk not identifying when a dilemma exists or making poor choices, since all aspects of the problem have not been identified.

BACKGROUND

Technology has moral properties, just as it has political properties (Brey 2000; Feenberg, 1999; Sclove, 1995; Winner, 1980). Winner (1980) argues that technological artifacts and systems function like laws, serving as frameworks for public order by constraining individuals' behaviors. Sclove (1995) argues that technologies possess the same kinds of structural effects as other elements of society, such as laws, dominant political and economic institutions, and systems of cultural beliefs. Data mining, being a technological artifact, is worthy of study from an ethical perspective due to its increasing importance in decision making, both in the private and public sectors. Computer systems often function less as background technologies and more as active constituents in shaping society (Brey, 2000). Data mining is no exception. Higher integration of data mining capabilities within applications ensures that this particular technological artifact will increasingly shape public and private policies.

Data miners and decision makers obviously are obligated to adhere to the law. But ethics are oftentimes more restrictive than what is called for by law. Ethics are standards of conduct that are agreed upon by cultures and organizations. Supreme Court Justice Potter Stewart defines the difference between ethics and laws as knowing the difference between what you have a right to do (legally, that is) and what is right to do. Sadly, a number of IS professionals either lack an awareness of what their company actually does with data and data mining results or purposely come to the conclusion that it is not their concern. They are enablers in the sense that they solve management's problems. What management does with that data or results is not their concern.

Most laws do not explicitly address data mining, although court cases are being brought to stop certain data mining practices. A federal court ruled that using data mining tools to search Internet sites for competitive information may be a crime under certain circumstances (Scott, 2002). In *EF Cultural Travel BV vs. Explorica Inc.*

(No. 01-2000 1st Cir. Dec. 17, 2001), the First Circuit Court of Appeals in Massachusetts held that Explorica, a tour operator for students, improperly obtained confidential information about how rival EF's Web site worked and used that information to write software that gleaned data about student tour prices from EF's Web site in order to undercut EF's prices (Scott, 2002). In this case, Explorica probably violated the federal Computer Fraud and Abuse Act (18 U.S.C. Sec. 1030). Hence, the source of the data is important when data mining.

Typically, with applied ethics, a morally controversial practice, such as how data mining impacts privacy, "is described and analyzed in descriptive terms, and finally moral principles and judgments are applied to it and moral deliberation takes place, resulting in a moral evaluation, and operationally, a set of policy recommendations" (Brey, 2000, p. 10). Applied ethics is adopted by most of the literature on computer ethics (Brey, 2000). Data mining may appear to be morally neutral, but appearances in this case are deceiving. This paper takes an applied perspective to the ethical dilemmas that arise from the application of data mining in specific circumstances as opposed to examining the technological artifacts (i.e., the specific software and how it generates inferences and predictions) used by data miners.

MAIN THRUST

Computer technology has redefined the boundary between public and private information, making much more information public. Privacy is the freedom granted to individuals to control their exposure to others. A customary distinction is between relational and informational privacy. Relational privacy is the control over one's person and one's personal environment, and concerns the freedom to be left alone without observation or interference by others. Informational privacy is one's control over personal information in the form of text, pictures, recordings, and so forth (Brey, 2000).

Technology cannot be separated from its uses. It is the ethical obligation of any information systems (IS) professional, through whatever means he or she finds out that the data that he or she has been asked to gather or mine is going to be used in an unethical way, to act in a socially and ethically responsible manner. This might mean nothing more than pointing out why such a use is unethical. In other cases, more extreme measures may be warranted. As data mining becomes more commonplace and as companies push for even greater profits and market share, ethical dilemmas will be increasingly encountered. Ten common blunders that a data miner may cause, resulting in potential ethical or possibly legal dilemmas, are (Skalak, 2001):

1. Selecting the wrong problem for data mining.
2. Ignoring what the sponsor thinks data mining is and what it can and cannot do.
3. Leaving insufficient time for data preparation.
4. Looking only at aggregated results, never at individual records.
5. Being nonchalant about keeping track of the mining procedure and results.
6. Ignoring suspicious findings in a haste to move on.
7. Running mining algorithms repeatedly without thinking hard enough about the next stages of the data analysis.
8. Believing everything you are told about the data.
9. Believing everything you are told about your own data mining analyses.
10. Measuring results differently from the way the sponsor will measure them.

These blunders are hidden ethical dilemmas faced by those who perform data mining. In the next subsections, sample ethical dilemmas raised with respect to the application of data mining results in the public sector are examined, followed briefly by those in the private sector.

Ethics of Data Mining in the Public Sector

Many times, the objective of data mining is to build a customer profile based on two types of data—factual (who the customer is) and transactional (what the customer does) (Adomavicius & Tuzhilin, 2001). Often, consumers object to transactional analysis. What follows are two examples; the first (identifying successful students) creates a profile based primarily on factual data, and the second (identifying criminals and terrorists) primarily on transactional.

Identifying Successful Students

Probably the most common and well-developed use of data mining is the attraction and retention of customers. At first, this sounds like an ethically neutral application. Why not apply the concept of students as customers to the academe? When students enter college, the transition from high school for many students is overwhelming, negatively impacting their academic performance. High school is a highly structured Monday-through-Friday schedule. College requires students to study at irregular hours that constantly change from week to week, depending on the workload at that particular point in the course. Course materials are covered at a faster pace; the duration of a single class period is longer; and subjects are often more difficult. Tackling the changes in a student's aca-