IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8928

Chapter IX

The Pitfalls of Knowledge Discovery in Databases and Data Mining

John Wang
Montclair State University, USA

Alan Oppenheim
Montclair State University, USA

ABSTRACT

Although Data Mining (DM) may often seem a highly effective tool for companies to be using in their business endeavors, there are a number of pitfalls and/or barriers that may impede these firms from properly budgeting for DM projects in the short term. This chapter indicates that the pitfalls of DM can be categorized into several distinct categories. We explore the issues of accessibility and usability, affordability and efficiency, scalability and adaptability, systematic patterns vs. sample-specific patterns, explanatory factors vs. random variables, segmentation vs. sampling, accuracy and cohesiveness, and standardization and verification. Finally, we present the technical challenges regarding the pitfalls of DM.

INTRODUCTION

"Knowledge discovery in databases (KDD) is a new, multidisciplinary field that focuses on the overall process of information discovery in large volumes of warehoused data" (Abramowicz & Zurada, 2001). Data mining (DM) involves searching through

This chapter appears in the book, *Data Mining: Opportunities and Challenges,* edited by John Wang.

databases (DBs) for correlations and/or other non-random patterns. DM has been used by statisticians, data analysts, and the management information systems community, while KDD has been mostly used by artificial intelligence and machine learning researchers. The practice of DM is becoming more common in many industries, especially in the light of recent trends toward globalization. This is particularly the case for major corporations who are realizing the importance of DM and how it can provide help with the rapid growth and change they are experiencing. Despite the large amount of data already in existence, much information has not been compiled and analyzed. With DM, existing data can be sorted and information utilized for maximum potential.

Although we fully recognize the importance of DM, another side of the same coin deserves our attention. There is a dark side of DM that many of us fail to recognize and without recognition of the pitfalls of DM, the data miner is prone to fall deep into traps. Peter Coy (1997) noted four pitfalls in DM. The first pitfall is that DM can produce "bogus correlations" and generate expensive misinterpretations if performed incorrectly. The second pitfall is allowing the computer to work long enough to find "evidence to support any preconception." The third pitfall is called "story-telling" and says "a finding makes more sense if there's a plausible theory for it. But a beguiling story can disguise weaknesses in the data." Coy's fourth pitfall is "using too many variables."

Other scholars have mentioned three disadvantages of mining a DB: the high knowledge requirement of the user; the choice of the DB; and the usage of too many variables during the process (Chen & Sakaguchi, 2000; Chung, 1999). "The more factors the computer considers, the more likely the program will find relationships, valid or not." (Sethi, 2001, p.69).

Our research indicated that the pitfalls of DM might be categorized into several groups. This chapter will first describe the potential roadblocks in an organization itself. Next, we explore the theoretical issues in contrast with statistical inference. Following that, we consider the data related issues that are the most serious concern. Here we find different problems related to the information used for conducting DM research. Then, we present the technical challenges regarding the pitfalls of DM. Finally, there are some social, ethical, and legal issues related to the use of DM— the most important of which is the privacy issue, a topic that is covered in Chapter 18.

ORGANIZATIONAL ISSUES

DM in an organization has both benefits and drawbacks. Naturally, the manner in which we interpret data will determine its ultimate benefit. Gathering data is generally not the issue here; there is much data already stored in data warehouses. We need to remember that DM, when misinterpreted, may lead to costly errors. There are a number of organizational factors and issues that also may be drawbacks and limit DM's implementation and effectiveness. These factors will be discussed in this section.

A recent survey of retail IT indicated that of the companies using DM, 53% attribute no direct benefit to their bottom line from DM. About 20% of respondents indicated that DM has contributed very little, while only 8.4% of the respondents indicated that DM has contributed substantially to profitability. Additionally, 64% of all companies