IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA ITB7505
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

Chapter 3

Benchmarking Data Mining Algorithms

Balaji Rajagopalan
Oakland University, USA

Ravi Krovi
University of Akron, USA

Data mining is the process of sifting through the mass of organizational (internal and external) data to identify patterns critical for decision support. Successful implementation of the data mining effort requires a careful assessment of the various tools and algorithms available. The basic premise of this study is that machine-learning algorithms, which are assumption free, should outperform their traditional counterparts when mining business databases. The objective of this study is to test this proposition by investigating the performance of the algorithms for several scenarios. The scenarios are based on simulations designed to reflect the extent to which typical statistical assumptions are violated in the business domain. The results of the computational experiments support the proposition that machine learning algorithms generally outperform their statistical counterparts under certain conditions. These can be used as prescriptive guidelines for the applicability of data mining techniques.

INTRODUCTION

The amount of data collected by businesses today is increasing at a phenomenal rate. Businesses face the challenge of integrating and correlating data related to online sales, offline sales, customer satisfaction surveys, and server log files. Data mining is the process of sifting through the mass of organizational (internal and external) data to identify patterns critical for decision support. Data mining techniques have been successfully employed in applications like fraud detection and

Previously Published in the *Journal of Database Management, vol.12, no.1,* Copyright © 2002, Idea Group Publishing.

This chapter appears in the book, Data Warehousing and Web Engineering by Shirley Becker.

bankruptcy prediction (Tam and Kiang, 1992; Lee, Han, and Kwon, 1996; Kumar, Krovi and Rajagopalan, 1997), strategic decision-making (Nazem and Shin, 1999) and database marketing (Brachman, 1996). Today, businesses have the unique opportunity for using such techniques for target marketing and customer retention. The analysis of this data is critical as more and more businesses use this information to analyze their competition, product or market. Intelligent tools which are based on rules derived from web mining can also play an important role in personalization related to site content and presentation. Recently, there has been considerable interest on how to integrate and mine such data (Mulvenna, Anand, & Buchner, 2000, Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, and Simoudis, 1996).

Business databases in general pose a unique problem for pattern extraction because of their complex nature. This complexity arises from anomalies such as discontinuity, noise, ambiguity, and incompleteness (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Analyzing such data is a key requirement for effective decision making. Decision support tools, however, vary in their ability to provide this degree of analytical processing. This is illustrated in Figure 1. Historically, decision makers had to manually deduce patterns using information generated by query reporting systems. One level of analytical sophistication above this was the ability to look at the data and perform analyses such as What-If and goal seeking. More recently, online analytical processing (OLAP) systems have shown promise in providing drill down capabilities. However, knowledge discovery of non-intuitive patterns is possible only by using data mining. These approaches can extend the power of decision support tools for more unstructured tasks.

Figure 1: A Framework for Analytical Processing