Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Which Companies are More Profitable When Using Data Mining Techniques?

Jerri Ligon
University of North Texas, ligonj@unt.edu

Jae-Sung Sim
University of North Texas, sim@unt.edu

Which Companies Are More Profitable When Using Data Mining Techniques?

Jae-Sung Sim (sim@unt.edu), Jerri Ligon (ligonj@unt.edu)
Department of Business Computer Information Systems, University of North Texas, Denton, Texas

Abstract

Over the past decade, many organizations have begun to routinely capture huge volumes of historical data describing their operations, products, and customers. The incredible amounts of data can be saved in electronic formats at reasonable costs with the development of latest data processing and storage technologies. The field of data mining addresses the question of how best to use this vast amount of historical data to discover general regularities and improve the process of making decisions. The concept of data mining is gaining acceptance by many companies as a means of seeking higher profits and lower costs. This research will review several important concepts of data mining and will survey many companies through questionnaires. After receiving survey results, the authors will analyze profits/costs effects by data mining while comparing these effects by SIC indexes.

Introduction

Over the past decade, many organizations have begun to routinely capture huge volumes of historical data describing their operations, products, and customers. In addition, today, the wide use of supermarket point-of-sale scanners, automatic teller machines, credit and debit cards, pay-per-view television, home shopping, electronic funds transfer, automated order processing, electronic ticketing, and the like makes collecting and processing massive data at an unprecedented rate possible. These incredible amounts of data can be saved in electronic formats in reasonable costs with the development of latest data processing and storage technologies.

People in this field have recognized that the effective use of data has tremendous potential and is a key element in the next generation of client-server enterprise information technology. However, until the boom of data mining techniques, many organizations have had difficulties for processing these ores into valuable information. An increasing awareness of data mining technology within many organizations and an attendant increase in efforts to capture, warehouse, and utilize historical data to support evidence-based decision making (Mitchell, 1999). The field of data mining addresses the question of how best to use this vast amount of historical data to discover general regularities and improve the process of making decisions.

The concept of data mining is gaining acceptance by many companies as a means of seeking higher profits and lower costs. Accordingly, many authors have investigated and written on the topic of data mining. While many authors have offered different views of the objectives, functions, and definitions of data mining, only a few[1] have researched the quantitative effects of data mining.

This research will review several important concepts of data mining and will survey many companies through questionnaires. After receiving survey results, the authors will analyze profits/costs effects of data mining while comparing these effects by SIC (Standard Industrial Classification) indexes.

Research Objective

The objective of this paper is to analyze profits/costs in each industry category when using data mining techniques, and thus to suggest data mining techniques to the suitable company categories.

Data Mining Concepts

Definition

Data mining is used today in a wide range of applications, such as tracking down criminals by the federal government, becoming an information broker from a supermarket, developing community knowledge for a business, cross selling, routing warranty claims, holding on to good customers, and weeding out bad customers (Berry and Linoff, 1997). Some applications include marketing, financial investment, fraud detection, manufacturing and production, and network management (Brachman et. al, 1996). Using data mining is not limited only to business fields. Data mining is also useful for data analysis of sky survey cataloging, planet mapping dataset of Venus, biosequence databases, and geoscience systems (Fayyad, Haussler, and Stolorz, 1996)

There are several terminologies similar to data mining. Data mining is a process that uses a variety of data analysis tools to discover meaningful patterns and relationships in data. Knowledge Discovery in Databases, also referred to as data mining, is the search for usable intelligence in large volumes of raw data. Data warehousing is the process of collecting and cleaning transactional data and making them available for on-line retrieval. There is a symbiotic relationship between data mining and data warehousing. Having a good data

[1] Use of a sophisticated data mining tool sold 3 times as many T-shirts in U.S. Open (Hoffman, 1999)

warehouse can facilitate effective data mining. Although data mining is possible without data warehouse, the data warehouse greatly improves the chances of success in data mining. (Inmon, 1996)

Steps for using data mining are: learning the application domain, creating a target dataset, data cleaning and preprocessing, data reduction and projection, choosing the function of data mining, choosing the data mining algorithm(s), data mining, interpretation, and using discovered knowledge (Fayyad, Piatetsky-Shapiro, and Smyth, 1996)

Reasons for Data Mining

The increasing interest in data mining follows from the confluence of several recent trends: the falling cost of large data storage devices and the increasing ease of collecting data over networks; the development of robust and efficient machine learning algorithms to process this data; and the falling cost of computational power and the enabling use of computationally intensive methods for data analysis (Mitchell, 1999)

Why they are using data mining? The reason is that data mining delivers value to industry. Data mining increases customer profitability, reduces customer/product churn, reduces costs through target marketing, uncovers new market, detects credit abuse and fraud, performs sales and trend analysis, and performs inventory management and control. Data mining tools simplify analyses by employing filters based upon user-defined criteria (Sauter, 1999)

Introduction-based data mining software uses machine-learning algorithms to analyze records in a firm's internal and customer databases, discovering patterns, transactional relationships, and rules that can predict future trends and indicate competitive opportunities. Raw data thus transformed is maintained in a data warehouse, providing support for a variety of analytical tasks and competitive decisions. As a result, questions that traditionally required extensive trial-and-error queries or statistical segmenting can be answered automatically. (Mena, 1996)

Hypotheses

This research will receive financial data for profits and costs for data mining from each company, and analyze these numbers to investigate following hypotheses.

H_1: There is a relationship between companies' profits with the use of data mining technique(s) and those without the use of data mining technique(s).

H_2: There is a relationship between companies' costs with the use of data mining technique(s) and those without the use of data mining technique(s).

H_3: There is a difference among companies' profits by company categories.

H_4: There is a difference among companies' costs by company categories.

Method

Participants

Approximately one hundred companies will be proportionally chosen by SIC indexes. The SIC system is a series of number codes that attempts to classify all business establishments by the types of products or services they make available. Establishments engaged in the same activity, whatever their size or type of ownership, are assigned the same SIC code. These definitions are considered important for standardization. Ten companies in each two digit category (e.g. 00, 10, 20, etc) in SIC indexes will be chosen. Each company will receive a survey questionnaire. In the event that a SIC category has a high disproportional response regarding use of data mining, additional surveys will be sent to companies in that category.

Materials

The survey contains several categorical and ratio questions:

- Demographic Information – name of company, title of interviewee, etc.
- The use of data mining technique(s) – Yes / No
- Purpose of using data mining
- Types of data mining techniques used
- Date of its first use
- Name(s) of Software
- Stage of data mining use
- The issues that affect the practical application of data mining.
- Profits
    - Before using data mining
    - After using data mining
- Costs
    - Before using data mining
    - After using data mining

Types of data mining techniques include market based analysis, memory-based reasoning, cluster detection, link analysis, decision trees and rule induction, artificial neural networks, genetic algorithms, and on-line analytic

processing (OLAP). The examples of data mining software include Blue Martini E-Merchandising System (Hibbard, 1999), Red Brick Data Mine (Richman, 1996), MineSet (Backhaus, 1999), Clementine, IMACS, MLC++, MOBAL, Recon, and etc. The examples of issues that affect its practical application in industry include insufficient training, inadequate tool support, data unavailability, overabundance of patterns, changing and time-oriented data, spatially oriented data, complex data types, and scalability (Brachman et. al, 1996).

Analysis

Profits and costs are dependent variables while the use of data mining and the types of data mining methods are independent variables. Two-way ANOVA test will be used to find the differences between profits/costs before and after using data mining methods.

Expected Results

The authors expect that the profits of companies are higher after using data mining techniques, and that the costs of companies are lower after using data mining techniques. The profits/costs for more customer-oriented companies are higher than less customer-oriented companies.

Future Research

Other quantitative effects should be investigated by more independent variables. The number of companies surveyed should be increased to more population (such as Fortune 500) for better reliability. Other industrial standard codes (e.g. NAICS) may be used to compare effects between SIC and other code.

Other issues from data mining can be researched in future. For example, privacy issues are further exacerbated now that the World Wide Web makes it easy for new data to be automatically collected and added to databases. As data mining tools and services become more widely available, privacy concerns are likely to intensify (Cranor, 1999). Data mining has the potential to equip companies with the ability to invade individual privacy. For example, information about a person's purchases and even which Internet sites they visit can be bought and sold without their knowledge or permission. (Wreden, 1997)

References

Backhaus, B. "Mining your engineering data," *CAE, Computer-Aided Engineering* 18(1), January 1999, pp. 56-59.

Berry, M. and Linoff, G. *Data Mining Techniques: for marketing, sales, and customer support.* John Wiley & Sons, Inc., New York, NY, 1997

Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. "Mining business databases," *Communications of the ACM* 39(11), November 1996, pp. 42-48.

Cranor, L. F. "Internet privacy," *Communications of the ACM* 42(2), February 1999, pp. 28-31.

Fayyad, U., Haussler, D. and Stolorz, P. "Mining scientific data," *Communications of he ACM* 39(11), November 1996, pp. 51-57.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM* 39(11), November 1996, pp. 27-34.

Hibbard, J. "Retail's human touch," *Informationweek* (726) March 22, 1999, pp. 79-80.

Hoffman, T. "U.S. Open online sales soar," *Computerworld* 33(37), September 13, 1999, p. 6.

Inmon, W H. "The data warehouse and data mining," *Communications of the ACM* 39(11), November 1996, pp. 49-50.

Mena, J. "Machine-learning the business: Using data mining for competitive intelligence," *Competitive Intelligence Review* 7(4), Winter 1996, pp. 18-25.

Mitchell, T.M. "Machine learning and data mining," *Communications of the ACM* 42(11) November 1999, pp. 30-36.

Richman, D. "Red Brick digs in to data mining market," *Computerworld* 30(23), June 3, 1996, p. 6.

Sauter, V.L. "Intuitive decision-making," *Communications of the ACM* 42(6), June 1999, pp. 109-115.

Wreden, N. "Insight or Intrusion? Data Mining's Effect on Privacy," *Communicationsweek* (650), February 17, 1997, p. 44.