



Chapter XI

Minimizing the Minus Sides of Mining Data

John Wang, Montclair State University, USA

Xiaohua Hu, Drexel University, USA

Dan Zhu, Iowa State University, USA

Abstract

This research explores the effectiveness of data mining in a commercial perspective. Statistical issues are specified first. Data accuracy and standardization follow. Diverse problems related to the information used for conducting a data mining research are identified. Also, the technical challenges and potential roadblocks in an organization itself are described. Certainly, minimizing the possible negative side of data mining, hopefully, without interfering its tremendous potentials, is a challenging task we are facing.

Introduction

Data mining (DM) or knowledge discovery in databases (KDD) is an experimental, exploratory, and iterative process that consists of a number of stages. “Knowledge discovery in databases is a new, multidisciplinary field that focuses on the overall process of information discovery in large volumes of warehoused data” (Abramowicz & Zurada, 2001; Pyle, 2003). DM involves searching through 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. Chen and Liu (2005) demonstrated that 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. Over the last 40 years, the tools and techniques to process structured information have continued to evolve from databases to data warehouses to data mining (DM). Although DM is still in its infancy, it is now being used in a wide range of industries and for a range of tasks in a variety of contexts (Marshall, McDonald, Chen, & Chung, 2004). There are several sectors that are more interested in DM: banking, medicine, insurance, retailing, and government (Firestone, 2005). The applications of DM are everywhere: from biomedical data (Hu & Xu, 2005) to mobile user data (Goh & Taniar, 2005), from data warehousing (Tjioe & Taniar, 2005) to intelligent Web personalization (Zhou, Cheung, & Fong, 2005), from analyzing clinical outcome (Hu et al., 2005) to mining crime patterns (Bagui, 2006), from mining geographical data (Savary, Gardarin, & Zeitouni, 2006) to an application on XML documents (Messaoud, Boussaid, & Rabaséda, 2006), from tapping the power of text mining (Fan, Wallace, Rich, & Zhang, 2006) to decision trees for mining data streams (Gama, Fernandes, & Rocha, 2006).

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 proving to fall deep into traps. Coy (1997) noted four pitfalls in DM. The first pitfall is if DM is performed incorrectly, it can produce “bogus correlations” and generate expensive misinterpretations. The second pitfall is allowing the computer to work long enough until it finds “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.” The fourth pitfall that Coy cautions is “using too many variables.”

24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/minimizing-minus-sides-mining-data/7520

Related Content

TLabel: A New OLAP Aggregation Operator in Text Cubes

Lamia Oukid, Omar Boussaid, Nadjia Benblidiaand Fadila Bentayeb (2016). *International Journal of Data Warehousing and Mining* (pp. 54-74).

www.irma-international.org/article/tlabel/171119

Scalable Biclustering Algorithm Considers the Presence or Absence of Properties

Abdelilah Balamane (2021). *International Journal of Data Warehousing and Mining* (pp. 39-56).

www.irma-international.org/article/scalable-biclustering-algorithm-considers-the-presence-or-absence-of-properties/272017

Multidimensional Design Methods for Data Warehousing

Oscar Romeroand Alberto Abelló (2011). *Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches* (pp. 78-105).

www.irma-international.org/chapter/multidimensional-design-methods-data-warehousing/53073

Enhancing the Process of Knowledge Discovery in Geographic Databases Using Geo-Ontologies

Vania Bogorny, Paulo Martins Engeland Luis Otavio Alavares (2008). *Data Mining with Ontologies: Implementations, Findings, and Frameworks* (pp. 160-181).

www.irma-international.org/chapter/enhancing-process-knowledge-discovery-geographic/7577

Data Mining Methods for Crude Oil Market Analysis and Forecast

Jue Wang, Wei Xu, Xun Zhang, Yejing Bao, Ye Pangand Shouyang Wang (2010). *Data Mining in Public and Private Sectors: Organizational and Government Applications* (pp. 184-203).

www.irma-international.org/chapter/data-mining-methods-crude-oil/44289