

Data Mining Fundamental Concepts and Critical Issues

John Wang

Montclair State University, USA

Qiyang Chen

Montclair State University, USA

James Yao

Montclair State University, USA

INTRODUCTION

Data mining is the process of extracting previously unknown information from large databases or data warehouses and using it to make crucial business decisions. Data mining tools find patterns in the data and infer rules from them. The extracted information can be used to form a prediction or classification model, identify relations between database records, or provide a summary of the databases being mined. Those patterns and rules can be used to guide decision making and forecast the effect of those decisions, and data mining can speed analysis by focusing attention on the most important variables.

BACKGROUND

We are drowning in data, but starving for knowledge. In recent years the amount or the volume of information has increased significantly. Some researchers suggest that the volume of information stored doubles every year. Disk storage per person (DSP) is a way to measure the growth in personal data. Edelstein (2003) estimated that the number has dramatically grown from 28MB in 1996 to 472MB in 2000.

Data mining seems to be the most promising solution for the dilemma of dealing with too much data having very little knowledge. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trend, patterns, exceptions and anomalies. The use of data mining can advance a company's position by creating a sustainable competitive advantage. Data

warehousing and mining is the science of managing and analyzing large datasets and discovering novel patterns (Davenport & Harris, 2007; Wang, 2006; Olafsson, 2006).

Data mining is taking off for several reasons: organizations are gathering more data about their businesses, the enormous drop in storage costs, competitive business pressures, a desire to leverage existing information technology investments, and the dramatic drop in the cost/performance ratio of computer systems. Another reason is the rise of data warehousing. In the past, it was often necessary to gather the data, cleanse it, and merge it. Now, in many cases, the data are already sitting in a data warehouse ready to be used.

Over the last 40 years, the tools and techniques to process data and information have continued to evolve from data bases to data warehousing and further to data mining. Data warehousing applications have become business-critical. Data mining can compress even more value out of these huge repositories of information. Data mining is a multidisciplinary field covering a lot of disciplines such as databases, statistics, artificial intelligence, pattern recognition, machine learning, information theory, control theory, operations research, information retrieval, data visualization, high-performance computing or parallel and distributed computing, etc (Zhou, 2003; (Hand, Mannila, & Smyth, 2001).

Certainly, many statistical models had emerged a long time ago. Machine learning has marked a milestone in the evolution of computer science. Although data mining 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 (Wang, 2003; Lavoie, Dempsey, & Connaway, 2006). Data mining is synonymous with knowledge discovery in databases, knowledge extrac-

tion, data/pattern analysis, data archeology, data dredging, data snooping, data fishing, information harvesting, and business intelligence (Han and Kamber, 2001).

MAIN FOCUS

Functionalities and Tasks

The common types of information that can be derived from data mining operations are associations, sequences, classifications, clusters, and forecasting. Associations happen when occurrences are linked in a single event. One of the most popular association applications deals with market basket analysis. This technique incorporates the use of frequency and probability functions to estimate the percentage chance of occurrences. Business strategists can leverage off of market basket analysis by applying such techniques as cross-selling and up-selling. In sequences, events are linked over time. This is particularly applicable in e-business for Website analysis.

Classification is probably the most common data mining activity today. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules from them. Clustering is related to classification, but differs in that no groups have yet been defined. Using clustering, the data-mining tool discovers different groupings within the data. The resulting groups or clusters help the end user make some sense out of vast amounts of data (Kudyba, & Hoptroff, 2001). All of these applications may involve predictions. The fifth application type, forecasting, is a different form of prediction. It estimates the future value of continuous variables based on patterns within the data.

Algorithms and Methodologies

Neural Networks

Also referred to as artificial intelligence (AI), neural networks utilize predictive algorithms. This technology has many similar characteristics to that of regression because the application generally examines historical data, and utilizes a functional form that best equates explanatory variables and the target variable in a man-

ner that minimizes the error between what the model had produced and what actually occurred in the past, and then applies this function to future data. Neural networks are a bit more complex as they incorporate intensive program architectures in attempting to identify linear, non-linear and patterned relationships in historical data.

Decision Trees

Megaputer (2006) mentioned that this method can be applied for solution of classification tasks only. As a result of applying this method to a training set, a hierarchical structure of classifying rules of the type “if...then...” is created. This structure has a form of a tree. In order to decide to which class an object or a situation should be assigned one has to answer questions located at the tree nodes, starting from the root. Following this procedure one eventually comes to one of the final nodes (called leaves), where the analyst finds a conclusion to which class the considered object should be assigned.

Genetic Algorithms (or Evolutionary Programming)

Genetic algorithms, biologically inspired search method, borrow mechanisms of inheritance to find solutions. Biological systems demonstrated flexibility, robustness and efficiency. Many biological systems are good at adapting to their environments. Some biological methods (such as reproduction, crossover and mutation) can be used as an approach to computer-based problem solving. An initial population of solutions is created randomly. Only a fixed number of candidate solutions are kept from one generation to the next. Those solutions that are less fit tend to die off, similar to the biological notion of “survival of the fittest”.

Regression Analysis

This technique involves specifying a functional form that best describes the relationship between explanatory, driving or independent variables and the target or dependent variable the decision maker is looking to explain. Business analysts typically utilize regression to identify the quantitative relationships that exist between variables and enable them to forecast into the future.

4 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/data-mining-fundamental-concepts-critical/10281

Related Content

The Nature of Intelligent Analytics

Zhaohao Sun and Andrew Stranieri (2021). *Intelligent Analytics With Advanced Multi-Industry Applications* (pp. 1-21).

www.irma-international.org/chapter/the-nature-of-intelligent-analytics/272776

Optimization of Hopfield Neural Network for Improved Pattern Recall and Storage Using Lyapunov Energy Function and Hamming Distance: MC-HNN

Jay Kant Pratap Singh Yadav, Zainul Abidin Jaffery and Laxman Singh (2022). *International Journal of Fuzzy System Applications* (pp. 1-25).

www.irma-international.org/article/optimization-of-hopfield-neural-network-for-improved-pattern-recall-and-storage-using-lyapunov-energy-function-and-hamming-distance/296592

Causal Machine Learning in Social Impact Assessment

Nuno Castro Lopes and Luís Cavique (2023). *Philosophy of Artificial Intelligence and Its Place in Society* (pp. 56-77).

www.irma-international.org/chapter/causal-machine-learning-in-social-impact-assessment/332600

Artificial Intelligence Techniques in Medicine and Health Care

Rezaul Begg (2008). *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 1750-1757).

www.irma-international.org/chapter/artificial-intelligence-techniques-medicine-health/24368

Rotational Invariance Using Gabor Convolution Neural Network and Color Space for Image Processing

Judy Gateri, Richard M. Rimiru and Michael Kimwele (2023). *International Journal of Ambient Computing and Intelligence* (pp. 1-11).

www.irma-international.org/article/rotational-invariance-using-gabor-convolution-neural-network-and-color-space-for-image-processing/323798