An analytic approach to select data mining for business decision

Jia-Lang Seng [a,*], T.C. Chen [b]

[a] Department & Graduate School of Accounting, College of Commerce, National Chengchi University, Taipei 115, Taiwan
[b] MICRONIX Technologies Inc., Taipei, Taiwan

ARTICLE INFO

ABSTRACT

Due to the information technology improvement and the growth of internet, enterprises are able to collect and to store huge amount of data. Using data mining technology to aid the data processing, information retrieval and knowledge generation process has become one of the critical missions to enterprise, so how to use data mining tools properly is user concern. Since not every user completely understand the theory of data mining, choosing the best solution from the functions which data mining tools provides is not easy. If user is not satisfied with the outcome of mining, communication with IT employees to adjust the software costs lots of time. To solve this problem, a selection model of data mining algorithms is proposed. By analyzing the content of business decision and application, user requirements will map to certain data mining category and algorithm. This method makes algorithm selection faster and reasonable to improve the efficiency of applying data mining tools to solve business problems.

© 2010 Published by Elsevier Ltd.

1. Introduction

1.1. Research background

Due to the information technology improvement and the growth of Internet, enterprises are able to collect and to store huge amount of data. People gradually realize that data is not equal to information that data should be further analyzed and extracted. Professionals are trained to analyze and interpret data, but the increases in data amount, data type, and analytical dimensions. Information technology has gone beyond storage, transmission, and processing. Data needs to be converted into information and knowledge in order to support decision making.

1.2. Research issue

Enterprises use data mining tools to support knowledge discovery and decision making. In this research, we develop a selection model to solve the research issue. This model recommends the most suitable algorithm after marketing professionals and analysts describe the business problems using a standard procedure and format. This model provides an algorithm standard as the foundation of dynamic data mining modeling.

This selection study of data mining mainly has two parts: the commercial problems analysis and the data mining algorithms analysis. Commercial problems analysis contains a general set of 22 problems. These problems relate to the banking, finance, insurance, telecommunication, retail, and manufacturing applications. They are classified into 12 business application categories according to their characteristics. Data mining algorithms analysis discuss five main types of data mining: association, classification, prediction, clustering, and profiling. In each category, several typical algorithms are listed and each of them is used to discover similar knowledge in different situations. The algorithm concepts, parameters, processes, and characters are compiled and compared to other algorithms in the same category.

1.3. Research limitation

Data mining has been applied to business area for more than two decades. There are countless application cases for different industries, different information requirement, and lots more circumstances. We adopt the literature survey to summarize approaches and concepts from our literature review to formulate the selection framework and produce the processing details into broader concepts and terms. This selection model focuses on matching the applicable data mining method with the characteristics of business decision and application.

2. Literature review

2.1. Data mining

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is

* Corresponding author.
  E-mail addresses: seng@nccu.edu.tw, jia.lang.seng@gmail.com (J.-L. Seng), jljan2@gmail.com (T.C. Chen).

how to make the database a competitive intelligence by converting seemingly meaningless data into useful information. How this challenge is met is critical because enterprises are increasingly relying on effective analysis of the information simply to remain competitive. By knowledge discovery in databases, interesting knowledge, regularities, and high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles. From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP). However, data mining goes far beyond the narrow scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques for data understanding (Han & Kamber, 2001). Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Choosing the data mining algorithm includes selecting method(s) to be used for searching for patterns in the data such as deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process. Data mining searches for patterns interest in a particular representational form set of such representations including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. The mining results which match the requirements will be interpreted and marshaled, to be taken into action or be presented to interested parties in the final step. The concept of data mining contains all activities and techniques utilizing the collected data to get implicit information and analyzing historical records to gain valuable knowledge.

2.1.1. Data mining method

Data mining methods refer to the function types that data mining tools provide. The conceptual definition of each data mining method and the assortment basis always differ for the ease of explanation, the consideration of present situation, or researcher background. Classification, association, prediction, clustering are usually the common methods in different works, while the term description, summarization, sequential rule may not always be used and listed in the first place.

Data mining methods include techniques which evolve from artificial intelligence, statistics, machine learning, OLAP and so on. These most often mentioned methods are classified into five categories according to their function types in business applications are shown in Table 1.

2.1.2. Data mining modeling

Data mining modeling is the critical part in developing business applications. Business applications, such as "cross-selling", will be turn into one or more of business problem and the goal of modeling is to formulate these business problems as data mining tasks.

Table 1
Data mining literature categories.

Author	Data mining categorization
Fayyad, Piatetsky-Shapiro, and Smyth (1996), Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996)	Classification, regression, clustering, summarization, dependency modeling, link analysis, and sequence analysis
Berry and Linoff (1997)	Classification, estimation, prediction, affinity grouping, clustering, and description

Fayyad, Piatetsky-Shapiro, and Smyth (1996) and Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996) state that data mining algorithms consists largely of some specific mix of three components:

- *The model:* There are two relevant factors. They are the function of the model and the representational form of the model. A model contains parameters that are to be determined from the data.
- *The preference criterion:* A basis for preference of one model or set of parameters over another, depending on the given data. The criterion is usually some form of goodness-of-fit function of the model to the data, perhaps tempered by a smoothing term to avoid over fitting, or generating a model with too many degrees of freedom to be constrained by the given data.
- *The search algorithm:* The specification of an algorithm for finding particular models and parameters, given data, a model, and a preference criterion.

In Berry and Linoff (1997), a mapping between business tasks and data mining techniques had been partially done. The choice of what data mining techniques to apply at a given point in the knowledge discovery processes depends on the particular data mining task to be accomplished and on the data available for analysis. The requirements of tasks dedicate to the functions of mining and the detail characteristics of tasks influence the feasibility between mining methods and business problems. The so called detail characteristics include data types, parameter varieties, hybrid approaches and so on. Slightly difference in the model will cause enormous performance changes. So the modeling stage affects the quality of data mining tools heavily (Table 2).

2.2. Business decision

Data mining assists many kinds of analysis works and the most popular business applications collected from the literature are listed as follows.

Finance and insurance

- Financial product cross-selling identifies additional types of products that customers could purchase which they currently are not purchasing. For example, a saving account customer would like to buy life insurance products of the original company.
- Telemarketing and direct mail marketing need marketing name list so data mining is in charge of figuring out the most responsive and valuable customers for certain products.
- Market segment analysis is one of the basic marketing researches where similar customers will be divides into same segmentations.
- Product mix analysis considers customer needs, marketing strategy and product line completeness to find out the best product mix.
- Credit card fraud detection and bad debt collection monitors and detects unusual activities of fraud and the occurrence of bad debt.
- Customer churn analysis is used to analysis the transaction record and behavior pattern to know the possibility of customer attrition, and then, enterprise will apply different recovery actions to customer base on their churn tendency and contribution to the company.
- Credit card application and personal loan rating classifies the customer into different level for loan or credit card applications.