Expert Systems with Applications 36 (2009) 4589–4594

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Decision analysis of data mining project based on Bayesian risk

Guangli Nie [a,b], Lingling Zhang [a,b,*], Ying Liu [b], Xiuyu Zheng [a,b], Yong Shi [b]

[a] School of Management, Graduate University of Chinese Academy of Sciences, Beijing 100080, China
[b] Research Center on Fictitious Economy and Data Science, CAS, Beijing 100080, China

ARTICLE INFO

*Keywords:*
Decision analysis
Data mining
Business intelligence

ABSTRACT

Data mining, an efficient method of business intelligence, is a process to extract knowledge from large scale data. As the augment of the size of enterprise and the data, data mining as a way to make use of the data become more and more necessary. But now most of the literatures only focus on the algorithm itself. Few literatures research what qualification to fulfill before the decision doing data mining from the perspective of the company manager. This paper discusses the factors affect the data mining project. Based on the Bayesian risk, we build a model taking the risk attitude of the top executive in account to help them make decision whether to do data mining or not.



## 1. Introduction

In recent years, data mining has been very well researched and a number of algorithms have been proposed (Hope & Korb, 2004; Wu, 1999; Ying, 2005) In order to prove the effectiveness of an algorithm, researchers test their algorithms in terms of accuracy, time cost and space cost, support, confidence, and lift are also measures of the interestingness of the rules or patterns in the databases (Wang, Strong, & Guarascio, 1994). For example, predictive accuracy is usually used to measure the effectiveness of classification learners, accepting a machine learner as superior to another if its predictive accuracy passes a statistical significance test (Hope & Korb, 2004).

However, researchers seldom consider the algorithm from the perspective of the company. Although some papers discuss the business issue, the aim is to prove the effectiveness of the algorithm (Strobel & Hrycej, 2006). Based on the limited resource and data quality, should my company use data mining techniques on our data? What could I get if I launch a data mining project? The current researchers cannot answer these questions.

Data mining is an application-driven technique(Chen & Liu, 2005; Sim, 2003). It has been widely used in many applications, from tracking criminals to brokering information for supermarkets, from developing community knowledge for a business to cross-selling, detecting the customer churn. Some applications include marketing, financial investment, fraud detection, manufacturing and production, and network management. Data mining is also useful for sky survey cataloging, mapping the datasets of Venus, biosequence databases, and Geosciences systems (Sim, 2003).

Although data mining are getting more widely used, the present research on data mining did not pay adequate attention to our "God", people who will use the algorithms. Research papers seldom help managers make decision whether to use data mining or not taking the risk, the finance condition, the effect brought to the company after the application of data mining, the data quality to account. In this paper, we try to build a mechanism to evaluate if a company is quantified to launch a data mining project. A score is used to measure whether the company should do data mining or not based on Bayesian risk.

The remaining of this paper is organized as follows. Section is the review on the application of data mining, the factors affecting data mining and the Bayesian analysis. Section 3 introduces how data quality affects data mining. Section 4 describes how to evaluate human factors and finance factors which are important to the success of a data mining project. Section 5 discusses importance of the support of the top executives. Bayesian Risk is presented in Section 6, followed by a case study in Section 7. Section 8 conclusions this paper and points out our further work.

## 2. Literature review

The literature review includes three parts. The first part is the review on data mining algorithms and the application of data mining. The second part is the factors that would affect the success of data mining project. The third part is the Bayesian risk analysis.

* Corresponding author. Address: School of Management, Graduate University of Chinese Academy of Sciences, Beijing 100080, China. Tel.: +86 10 82680396; fax: +86 10 82680698.
*E-mail addresses:* sdungl@163.com (G. Nie), zhangll@gucas.ac.cn (L. Zhang), yingliu@gucas.ac.cn (Y. Liu), zhengxiuyu06@mails.gucas.ac.cn (X. Zheng), yshi@gucas.ac.cn (Y. Shi).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2008.05.014

### 2.1. Data mining application

Data mining models can be obtained by employing a lot of algorithms. The two most common supervised modeling methods are classification and regression (Moreno Garcı'a et al., 2008).

Many unsupervised algorithms are proposed such as association rule mining, clustering (Jain, Murty, & Flynn, 1999), sequence mining and so on. Association rule mining was first introduced by Agrawal et al. in the context of transaction databases (Agrawal, Imielinski, & Swami, 1993a, 1993b). Data clustering has been studied in the Statistics (Dubes & Jaiu, 1980; Duda & Hart, 1973; Lee, 1981; Murtagh, 1983), Machine Learning (Peter, James, & Matthew, 1988) and Database (Raymond & Han, 1994) communities with different, methods and different emphases.

A lot of papers focus on the improvement of the existing algorithms such as naive Bayes, association rule mining (Chen, Liu, Yu, Wei, & Zhang, 2006; Chen et al., 2007). The frequent association mining problem has drawn much attention over the past decade. Many algorithms have been proposed to improve the mining of frequent itemsets (Han, Pei, & Yin, 2000). The class imbalance problem is an important issue in classification of Data mining. An approach was proposed to mining the multi-relational imbalanced database (Lee et al., 2007).

It is the application that drives the development of data mining algorithms. As the development of the information technology, it is quite easy for organization to collect data at a very low cost especially the organization offering service, the banking for example. The traditional tools used to analyze the data cannot meet the need of the enterprise to process so much data. Many people both from academic and industrial generally realize the importance of the data mining and made some effort on it.

Shu-Hsien Liao reviewed the literatures from 1995 to 2002 and summarized the application of data mining as follows: cross-sales, deviation detection, finance, organizational learning, user-guided query construction, interface, consumer behaviors/service, semantic indexing, data quality, health care management, knowledge refinement, prediction of failure, marketing, software integration, knowledge warehouse, grid services, hypermedia (Liao, 2003).

The use of data mining and visualization techniques for decision support in planning and regional level management of Slovenian public health-care has been demonstrated applicable and practicable (Lavrac et al., 2007). Financial markets generate large volumes of data. Nearly 10 kinds of application of data mining for finance were listed such as portfolio management, probability distribution estimation for financial data, stock charting by Hui Wang et al. (Hui Wang & Andreas S. Weigend).

Data mining is frequently adopted to discover the valuable information from the huge database (Chen & Lin, 2007). In data mining, association rule mining is widely applied to market basket analysis or transaction data analysis (Agrawal, Imielinski & Swami, 1993; Srikant and Agrawal, 1997). An approach based on association rule mining was proposed to do product assortment and shelf space allocation. Due to the increasingly massive amount of biology, chemistry and clinical data, data mining was used to improve drug delivery after mining public and/or proprietary data (Ekins et al., 2006). Data mining was also used in project management. A process to refine association rules, based on the generation of unexpected patterns, is proposed. The goal is to generate strong association rules between attributes that can be obtained early in the project and the final software size (Moreno Garci'a et al., 2004). Data mining can play an important role in Marketing, especially customers relationship management (CRM), the CRM of churn email for example. Data mining also work in Internet Service Providers. Data mining (decision tree and support vector machines algorithms) can also be used for making cancer predictions with Analysis of gene expression data (Shah & Kusiak, 2007). Data Mining classification techniques was used in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS (Kirkos et al., 2007). Data mining technology is introduced to fault diagnosis of rotating machinery, and a new method based on C4.5 decision tree and principal component analysis (PCA) was proposed (Sun, Chen, & Li, 2007). Data mining was used in the mental health (Diederich et al., 2007). A deployed data mining application system for Motorola whose intended use was for identifying causes of cellular phone failures was developed.

In academic area, data mining work is a good tool. Data mining was used to mine the chemical information in GRID environment (Maran, Sild, Kahn, & Takkis, 2007).

What all these researches focus on is how to design a new algorithm or to improve the existing algorithms to mine more quickly and accurately. The application researches are based on assumption that the real situation is suitable to do data mining. All these research works rely on a hypothesis that the data they mine is suitable for data mining. They failed to take account of the real status of the organization where the algorithms would be implemented.

### 2.2. The factor affecting the data mining project

Not like the massive papers concerning algorithms, little attention has been paid on the analysis about the data mining project. There are few literatures about the success of data mining project. Although the data quality was once discussed, the main aim of the discussion was how to improve the data quality. The problem falls into two categories. The first category is related to inconsistency among systems such as format, syntax and semantic inconsistencies. The second category is related to inconsistency with reality as it is exemplified by missing, obsolete and incorrect data values and outliers. The usefulness of the results produced by data mining methods can be critically impaired by several factors such as (1) low quality of data, including errors due to contamination, or incompleteness due to limited bandwidth for data acquisition, and (2) inadequacy of the data model for capturing complex probabilistic relationships in data.

The cost of the data mining including the payment on the project and the cost because of the misjudgment of the data mining model attracts some attention. Some research shows that factors such as the probability of intrusion and the costs of responding to detected intrusions must be taken into account in order to compare the effectiveness of machine learning algorithms over the intrusion detection domain.

There are some works on the fundamental trade-off in distributed data mining; namely, the trade-off between the efficiency and cost-effectiveness of a distributed data mining application on one side, and the accuracy and reliability of the resulting predictive system on the other side.

To the best of our knowledge, there is seldom research on the factor that will affect the success of the data mining project such as the human factor, finance factors, the support of the chief executives, risk attitude of the executive. All of the factors listed above would be discussed separately in Sections 3 and 4 in this paper.

### 2.3. The Bayesian analysis

The research on Bayesian analysis is quite mature now. There are a lot of research achievements on this topic. Assuming that the uncertainties involved in the decision problem can be considered to be unknown numerical quantities, and will represent them by $\theta$ (a vector or matrix) which is commonly called the state of nature, $\theta$ will be used to denote the set of all possible states of nature. Typically some experiments can be conducted to obtain statistical information about them. In decision theory, an attempt is made to