ELSEVIER

Information & Management 37 (2000) 271–281

INFORMATION & MANAGEMENT

www.elsevier.com/locate/dsw

Briefings

Methodological and practical aspects of data mining

A. Feelders[a,*], H. Daniels[a,b], M. Holsheimer[c]

[a]*Department of Economics and Business Administration, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands*

[b]*Rotterdam School of Management, Institute of Advanced Management Studies, PO Box 1738, 3000 DR Rotterdam, The Netherlands*

[c]*Data Distilleries, Kruislaan 419, 1098 VA Amsterdam, The Netherlands*

Received 15 October 1998; accepted 5 September 1999

Abstract

We describe the different stages in the data mining process and discuss some pitfalls and guidelines to circumvent them. Despite the predominant attention on analysis, data selection and pre-processing are the most time-consuming activities, and have a substantial influence on ultimate success. Successful data mining projects require the involvement of expertise in data mining, company data, and the subject area concerned. Despite the attractive suggestion of 'fully automatic' data analysis, knowledge of the processes behind the data remains indispensable in avoiding the many pitfalls of data mining.

Keywords: Data mining; Knowledge discovery in databases; Data quality

1. Introduction

Data mining is receiving more and more attention from the business community, as witnessed by frequent publications in the popular IT-press, and the growing number of tools appearing on the market. The commercial interest in data mining is mainly due to increasing awareness of companies that the vast amounts of data collected on customers and their behavior contain valuable information. If the hidden information can be made explicit, it can be used to improve vital business processes. Such developments are accompanied by the construction of data warehouses and data marts. These are integrated databases that are specifically created for the purpose of analysis rather than to support daily business transactions.

Many publications on data mining discuss the construction or application of algorithms to extract knowledge from data. The emphasis is generally on the analysis phase. When a data mining project is performed in an organizational setting, one discovers that there are other important activities in the process. These activities are often more time consuming and have an equally large influence on the ultimate success of the project.

Data mining is a multi-disciplinary field, that is at the intersection of statistics, machine learning, database management, and data visualization. A natural question comes to mind: to what extent does it provide a new perspective on data analysis? This question has

* Corresponding author. Tel.: +31-134668201; fax: +31-134663377.
E-mail address: a.j.feelders@kub.nl (A. Feelders)

0378-7206/00/$ – see front matter
PII: S0378-7206(99)00051-8

received some attention within the community. A popular answer is that data mining is concerned with the extraction of knowledge from *really large* data sets. In our view, this is not the complete answer. Company databases indeed are often quite large, especially if one considers data on customer transactions. One should however take into account the fact that:

- Once the data mining question is specified accurately, only a small part of this large and heterogeneous database is of interest.
- Even if the remaining dataset is large, a sample often suffices to construct accurate models.

If not necessarily in the size of the dataset, where does the contribution of the data mining perspective lie? Four aspects are of particular interest:

1. There is a growing need for valid methods that cover the *whole* process (also called Knowledge Discovery in Databases or KDD), from problem formulation to the implementation of actions and monitoring of models. Methods are needed to identify the important steps, and indicate the required expertise and tools. Such methods are required to improve the quality and controllability of the process.
2. If it is going to be used on a daily basis within organizations, then a better integration with existing information systems infrastructures is required. It is, for example, important to couple analysis tools with Data Warehouses and to integrate data mining functionality with end-user software, such as marketing campaign schedulers.
3. From a statistical viewpoint it is often of dubious value because of the absence of a *study design*. Since the data were not collected with any set of analysis questions in mind, they were not sampled from a pre-defined population, and data quality may be insufficient for analysis requirements. These anomalies in data sets require a study of problems related with analysis of 'non-random' samples, data pollution, and missing data.
4. Ease of interpretation is often understood to be a defining characteristic of data mining techniques. The demand for explainable models leads to a preference for techniques such as rule induction, classification trees, and, more recently, bayesian networks. Furthermore, explainable models encourage the explicit involvement of domain experts in the analysis process.

2. Required expertise

Succesful data mining projects require a *collaborative* effort in a number of areas of expertise.

2.1. Subject area expertise

Scenarios, in which the subject area expert provides the data analyst with a dataset and a question, expecting the data analyst to return 'the answer', are doomed to fail. The same is true for situations where the subject area expert does not have a specific question and ask the analyst to come up with some interesting results. Data mining is not some 'syntactical exercise', but requires knowledge of the *processes behind the data*, in order to:

- Determine useful questions for analysis;
- Select potentially relevant data to answer these questions;
- Help with the construction of useful features from the raw data; and
- Interpret (intermediate) results of the analysis, suggesting possible courses of action.

2.2. Data expertise

Knowledge of available data within — and possibly outside — the organization is of primary importance for the selection and pre-processing of data. The data expert knows exactly where the data can be found, and how different data sources can be combined. Peculiarities of data conversions that took place years ago can have substantial influence on the interpretation of results; for example:

> A large insurance company takes over a small competitor. The insurance policy databases are joined, but the start-date of the policies of the small company are set equal to the conversion date, because only the most recent mutation date was recorded by the small company. Lacking the knowledge of this conversion, one might believe there was an enormous 'sales peak' in the year of conversion.