Liet. matem. rink., T. 43, spec. nr., 248–253

Data mining approach on information detection using intelligent miner for data

Regina KULVIETIENĖ, Jelena MAMČENKO (VGTU)
e-mail: regina_kulvietiene@gama.vtu.lt, jelena@gama.vtu.lt

Introduction

Data Mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases. The genesis of the field came with the realization that traditional decision – support methodologies, which combine simple statistical techniques with executive information systems, do not scale to the point where they can deal with large databases and data warehouses within the time limits imposed by today's business environment. Data Mining has captured the imagination of the business academic worlds, moving very quickly from a niche research discipline in the mid – eighties to a flourishing field today. In fact, 80% of the Fortune 500 companies are currently involved in a data mining pilot project or have already deployed one or more data mining production systems.

In this article is presented to Data Mining technique and its application. This technique provides the tools, which can help to discover new contexts and hence new things about data. Mining any business will enable to make decisions based upon real knowledge.

1. Data mining

Data mining is about discovering new things about business from the collected data. In reality what we are normally doing is making a hypothesis about the business issue that is addressing and then attempting to prove or disprove hypothesis by looking for data to support or contradict the hypothesis [1].

For example, suppose that as a retailer, believe that customers from "out of town" visit larger inner city stores less often than other customers, but when they do so they make larger purchases. To answer this type of question it can be simply formulate a database query looking, for example, at the branches, their locations, sales figures, customers and then compile the necessary information (average spend per visit for each customer) to prove this hypotheses. However, the answer discovered might only be true for a small highly profitable group of out-of-town shoppers who visited inner-city stores at the weekend. At the same time, out-of-town customers (perhaps commuters) may visit the store during the week and spend exactly the same way as other customers. In this case, initial hypothesis test may indicate that there is no difference between out-of-town and inner-city shoppers.

Data mining uses an alternative approach beginning with the premise that we do not know what patterns of customer behaviors exist. In this case it might be simply asked the question, what are their relationships (sometimes use the term *correlations*) between what my customers spend and where they come from? This should include the out-of-town, weekend shopper. Data mining therefore provides answers, without having to ask specific questions. The difference between the two approaches is summarized in Fig. 1 [1].

It is difficult to make definitive statements about an evolving area – and surely data mining is an area in very quick evolution. However, it needs a framework within which to position and better understand the subject.

Although there is no one single definition of data mining that would meet with universal approval, the following definition is generally acceptable [2, 3]: "Data Mining is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions." The highlighted words in the definition lend insight into the essential nature of data mining and help to explain the fundamental differences between it and the traditional approaches to data analysis such as query and reporting and online analytical processing (OLAP). In essence, data mining is distinguished by the fact that it is aimed at the discovery of information, without a previously formulated hypothesis [1, 6].

First, the information discovered must have been previously unknown [3]. Although this sounds obvious, the real issue here is that it must be unlikely that the information could have been hypothesized in advance; that is, the data miner is looking for something that is not intuitive or, perhaps, even counterintuitive. The further away the information is from being obvious, potentially the more value it has. Data mining can uncover information that could not even have been hypothesized with other approaches.

Second, the new information must be valid [3] This element of the definition relates to the problem of overoptimism in data mining; that is, if data miners look hard enough in a large collection of data, they are bound to find something of interest sooner or later. For example, the potential number of associations between items in customers' shopping baskets rises exponentially with the number of items. The possibility of spurious results applies to all data mining and highlights the constant need for post-data-mining validation and sanity checking.

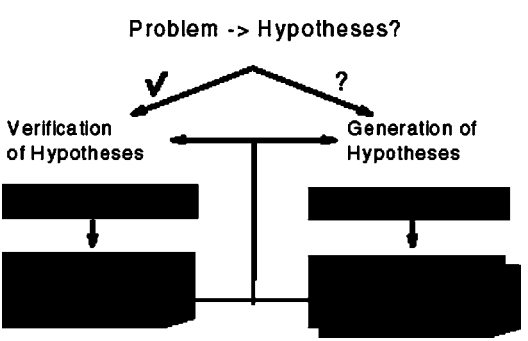

Fig. 1. Standard and data mining approach on information detection.

Third, and most critically, the new information must be actionable, that is, it must be possible to translate it into some business advantage [3]. In the case of the classic example of the retail store manager, who, using data mining, discovered that there was a strong association between the sales of diapers and beer on Friday evenings, clearly he could leverage the results of the analysis by placing the beer and diapers closer together in the store or by ensuring that the two items were not discounted at the same time. In many cases, however, the actionable criterion is not so simple.

The ability to use the mined data to inform crucial business decisions is another critical environmental condition for successful commercial data mining and underpins data mining's strong association with and applicability to business problems [5].

1.1. *Types of techniques*

In general, data mining techniques can be divided into two broad categories [6, 7, 8]:

Discovery data mining

Discovery data mining is applied to a range of techniques, which find patterns inside data without any prior knowledge of what patterns exist. The following are examples of discovery mining techniques:

- clustering,
- link analysis,
- frequency analysis.

Predictive Mining

Predictive data mining is applied to a range of techniques that find relationships between a specific variable (called the *target variable*) and the other variables in data. The following are examples of predictive mining techniques:

- classification,
- value prediction.

2. Data mining and business intelligence

We use business intelligence as a global term for all the processes, techniques, and tools that support business decision-making based on information technology. The approaches can range from a simple spreadsheet to a major competitive intelligence undertaking. Data mining is an important new component of business intelligence [1].

In general, the value of the information to support decision-making increases from the bottom of the pyramid to the top. A decision based on data in the lower layers, where there are typically millions of data records, will typically affect only a single customer transaction. A decision based on the highly summarized data in the upper layers is much more likely to be about company or department initiatives or even major redirection. Therefore we generally also find different types of users on the different layers. A database administrator works primarily with databases on the data source and data warehouse level, whereas business analysts and executives work primarily on the higher levels of the pyramid.

3. Introduction to the Intelligent Miner

The IBM Intelligent Miner for Data (IM in this article) is leading the way in helping customers identify and extract high-value business intelligence from their data assets [6]. The process is one of discovery. Companies are empowered to leverage information hidden within enterprise data and discover associations, patterns and trends; detect deviations; group and classify information; and develop predictive models. IBM's award winning Intelligent Miner was released in 1996.

It enables users to mine structured data stored inconventional databases or flat files. Customers and partners have successfully deployed its mining algorithms to address such business areas as market analysis, fraud and abuse, and customer relationship management [6].

The Intelligent Miner offerings are intended for use by data analysts and business technologists in areas such as marketing, finance, product management, and customer relationship management. In addition, the text mining technologies have applicability to a wide range of users who regularly review or research documents – for example, patent attorneys, corporate librarians, public relations teams, researchers, and students.

The IBM Intelligent Miner is a suite of statistical, processing, and mining functions that can be used to analyze large databases. It also provides visualization tools for viewing and interpreting mining results.

The Intelligent Miner provides a complete graphical user interface with TaskGuides that leads through the steps of creating the different Intelligent Miner objects. General help for each TaskGuide provides additional information, examples, and valid values for the controls on each page. In the sections that follow we introduce the data mining technology and the data mining process of the Intelligent Miner. We also explain in general the statistical, processing, and mining functions that Intelligent Miner provides.

4. Data Mining with the Intelligent Miner

Data mining is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data [9]. It can be used to extract information to form a prediction or classification model, or to identify similarities between database records. The resulting information can help to make more informed decisions. The Intelligent Miner helps organizations perform data mining tasks. For example, a retail store might use the Intelligent Miner to identify groups of customers that are most likely to respond to new products and services or to identify new opportunities for cross selling. An insurance company might use the Intelligent Miner with claims data to isolate likely fraud indicators.

5. DB2 Intelligent Miner Scoring

IM Scoring describes the supported mining functions, the scoring process, and the scoring functions and data types involved. It is also covers PMML and the model conversion

facility (standard format for data mining models) [4].

IM Scoring is an add-on service to DB2 that extends the capabilities of DB2 to include data mining functions. Using the IM Scoring functions, we can import certain types of mining models into a DB2 table and apply the models to data within DB2.

The results of applying the model are referred to as scoring results. They differ in content according to the type of model applied.

It extends database capabilities and enables users to deploy data mining analytics in real-time. It can be possible to employ these in business intelligence and operational applications to better serve business and consumer users alike – by providing more informed recommendations, more personalized treatment of business and consumer customers, or by continual model improvement in response to a specific situation [4].

Conclusions

The Data Mining technology helps to identify and extract high-value business intelligence from data assets. Through a process of ”knowledge discovery”, organization can leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to real competitive advantage.

IBM DB2 Intelligent Miner Scoring is built as an extender to DB2, it works directly from the relational database, and speeds up the data mining process, resulting in the ability to make quicker decisions from a host of culled data. IBM DB2 Intelligent Miner Scoring is also compatible with Oracle databases. And it is also allows businesses to make faster decisions and enhance relationships with their customers by driving data mining intelligence to key customer, supplier, and employee information.

Presently in Vilnius Gediminas Technical University Data Mining technique is widespread. One of the application appliances was taken to find some reliance on data set taken from telecommunication provider and bank information system. The main target was concentrated on customers' characteristics discovering using telecom data and data from bank information system. In both cases demographic clustering technique was applied. All customers were distributed to groups, which were interpreted as special offer for one or another group. Results were interpreted from business perspective. This test was successfully applied.

References

[1] C. Baragoin, Ch.M. Andersen, S. Bayerl, *IBM Redbooks, Mining your Own Business in Telecom*, ITSO, IBM Corp. (2001).

[2] В. Дюк, А. Самойленко, *Data Mining учебный курс*, Санкт-Петербург (2001).

[3] *IBM's Data Mining Technology*, White paper, Data Management Solution, First edition, April (1996), pp. 1–11.

[4] *IBM Intelligent Miner Scoring*, Administration and Programming for DB2.

[5] *Knowledge Discovery Through Data Mining: What is Knowledge Discovery*? Tandem Computers Inc. (1996).

[6] R. Kulvietienė, J. Mamčenko, Data Mining Technologies based on IBM Intelligent Miner, *Liet. matem. rink.*, **42** (spec. Nr.), 621–629 (2002).
[7] R. Kulvietienė, J. Mamčenko, Data Mining techniques, in: *Informacinės technologijos*: *teorija, praktika, inovacijos*, Respublikinės mokslinės konferencijos medžiaga, Alytus (2003), pp. 56–63.
[8] R.S. Michalski, K.A. Kaufman, Data mining and knowledge discovery: a review of issues and multistrategy approach, in: *Machine Learning and Data Mining: Methods and Applications*, West Sussex, England (1998), pp. 71–105.
G. Piatetsky–Shapiro, The data mining industry coming of age, *IEEE Intelligent Systems*, November/December, 32–34 (1999).

Duomenų gavybos metodas informacijos radimui, naudojant intelektualią duomenų paiešką

R. Kulvietienė, J. Mamčenko

DB2 Intelligent Miner Scoring praplėčia duomenų bazių galimybes ir leidžia vartotojams iš karto išskleisti duomenų gavybos loginės analizės techniką.

Duomenų gavybos technologijų sistemos realizuoja naują duomenų analizės formą, paremtą intelektualiais sprendimais, leidžia gauti iš duomenų bazės žymiai gilesnes žinias, negu sudėtingiausios užklausos ir iš jų suformuotos ataskaitos.