DATA MINING PROLOGUES

K.Sankar Lecturer / M.E., (P.hD).,
D.V.Rajkumar M.C.A., M.Phil Lecturer

Abstract

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Introduction

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a *difference of kind rather than degree*. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a customer who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.

Reflecting this conceptualization of data mining, some observers consider data
mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation.

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source.

In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing it have also played a role. Over the past several years there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year. At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.

Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. However, the proliferation of data mining has raised some implementation and oversight issues as well. These include concerns about the quality of the data being analyzed, the interoperability of the databases and software between agencies, and potential infringements on privacy. Also, there are some concerns that the limitations of data mining are being overlooked as agencies work to emphasize their homeland security initiatives.

Limitations of Data Mining

While data mining products can be very powerful tools, they are not self sufficient

applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related.

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use.

However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations).

Data Mining Uses

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.

Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor. In the public sector, data mining applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent Medicare payments. The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs.

Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures. Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. Two initiatives that have attracted significant attention include the now-discontinued Terrorism Information Awareness (TIA) project conducted by the Defense Advanced Research Projects Agency (DARPA), and the now-canceled Computer-Assisted Passenger Prescreening System II (CAPPS II) that was being developed by the Transportation Security Administration (TSA). CAPPS II is being replaced by a new program called Secure Flight.

Data Mining Issues

As data mining initiatives continue to evolve, there are several issues Congress may decide to consider related to implementation and oversight. These issues include, but are not limited to, data quality, interoperability, mission creep, and privacy. As with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project's outcome.

Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the

lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data.

To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database (e.g., ensuring that "no" is represented as a 0 throughout the database, and not sometimes as a 0, sometimes as a N, etc.), accounting for missing data points, removing unneeded data fields, identifying anomalous data points (e.g., an individual whose age is shown as 142 years), and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY).

Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing through e-government and homeland security initiatives. For data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously, and to help ensure the compatibility of data mining activities of different agencies. Data mining projects that are trying to take advantage of existing legacy databases or that are initiating first-time collaborative efforts with other agencies or levels of government (e.g., police departments in different states) may experience interoperability problems. Similarly, as agencies move forward with the creation of new databases and information sharing efforts, they will need to address interoperability issues during their planning stages to better ensure the effectiveness of their data mining projects.

Mission Creep

Mission creep is one of the leading risks of data mining cited by civil libertarians, and represents how control over one's information can be a tenuous proposition. Mission creep refers to the use of data for purposes other than that for which the data was originally collected. This can occur regardless of whether the data was provided voluntarily by the individual or was collected through other means. Efforts to fight terrorism can, at times, take on an acute sense of urgency. This urgency can create pressure on both data holders and officials who access the data. To leave an available resource unused may appear to some as being negligent.

Data holders may feel obligated to make any information available that could be used to prevent a future attack or track a known terrorist. Similarly, government officials responsible for ensuring the safety of others may be pressured to use and/or combine existing databases to identify potential threats. Unlike physical searches, or the detention of individuals, accessing information for purposes other than originally intended may appear to be a victimless or harmless exercise. However, such information use can lead to unintended outcomes and produce misleading results. One of the primary reasons for misleading results is inaccurate data. All data collection efforts suffer accuracy concerns to some degree. Ensuring the accuracy of information can require costly protocols that may not be cost effective if the data is not of inherently high economic value.

In well-managed data mining projects, the original data collecting organization is likely to be aware of the data's limitations and account for these limitations accordingly. However, such awareness may not be communicated or heeded when data is used for other purposes. For example, the accuracy of information collected through a shopper's club card may suffer for a variety of reasons, including the lack of identity authentication when a card is issued, cashiers using their own cards for customers who do not have one, and/or customers who use multiple cards. For the purposes of marketing to consumers, the impact of these inaccuracies is negligible to the individual. If a government agency were to use that information to target individuals based on food purchases associated with particular religious observances though, an outcome based on inaccurate information could be, at the least, a waste of resources by the government agency, and an unpleasant experience for the misidentified individual.

As the March 2004 TAPAC report observes, the potential wide reuse of data suggests that concerns about mission creep can extend beyond privacy to the protection of civil rights in the event that information is used for "targeting an individual solely on the basis of religion or expression, or using information in a way that would violate the constitutional guarantee against self-incrimination."

Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes (mission creep). For example, some experts suggest that anti-terrorism data mining applications might also be useful for combating other types of crime as well.

So far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that trade offs may need to be made regarding privacy to ensure security. Other observers suggest that existing laws and regulations regarding

privacy protections are adequate, and that these initiatives do not pose any threats to privacy. Still other observers argue that not enough is known about how data mining projects will be carried out, and that greater oversight is needed.

There is also some disagreement over how privacy concerns should be addressed. Some observers suggest that technical solutions are adequate. In contrast, some privacy advocates argue in favor of creating clearer policies and exercising stronger oversight. As data mining efforts move forward, Congress may consider a variety of questions including, the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and the possible application of the Privacy Act to these initiatives.

Data mining as a Business

On the basis of these issues it is also clear under what conditions data mining as a business can be successful. Data mining companies vary considerably, but in general a company either sells consultancy, tools or a combination of the two. If one takes point four of the previous section into account then it is clear that a vendor that only offers horizontal data mining tools will always be in competition with vendors of data base management environments, OLAP, reporting and query tools. They have inherently a bigger potential market then data mining tool vendors. For vendors of horizontal database management systems and query tools it is relatively easy to enhance their product with data mining capabilities or to buy a small innovative data mining company in order to get access to data mining expertise.

Furthermore, the data mining vendor is probably the last one to enter the client's site and will find a database environment that is already up and running. This is one of the main reasons that it is almost impossible for a company to survive on the basis of sales of horizontal data mining tools only. There is no room in the market for independent vendors of horizontal data mining tools. In this light there are a number of strategies that a data mining company can follow: find a vertical market and specialize, sell to a strong vendor of horizontal solutions, or simply quit the business.

In selling data mining consultancy, it is not easy to find a market that sustains a healthy business in the long run. The problem is that data mining per se deals with finding deep knowledge that is specific to an organization. Also, as Kohavi et al. (2004) observe, every problem is different. The client usually knows his business better than the consultant. Only by building up specific expertise concerning the application of data mining techniques in a vertical segment the data mining business can survive.

References

Adriaans, P. (2002). Backgrounds and general trends. In J. Meij (Ed.), *Dealing with the dataflood, mining data,text and multimedia* (pp. 16–25). STT

Beweton, The Hague, Netherlands. Adriaans, P. (2002a). Production control. In W. Kl¨osgen, & J. M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery*. Oxford University Press.

Adriaans, P., & Zantinge, D. (1996). *Data mining*. Addison-Wesley

BIBLIOGRAPHIES

K.Sankar M.E., (P.hD)., / Lecturer
K.S.R College of Engineering
Tiruchengode, Namakkal
Tamil Nadu, India
Email : san_kri_78@rediffmail.com

D.V.Rajkumar M.C.A., M.Phil/ Lecturer
K.S.R College of Engineering
Tiruchengode, Namakkal
Tamil Nadu, India dvrajkumarmca@rediffmail.com