A NEW SIGHT INTO DATA MINING

Er. Satinderpal Singh*, Er.Sheilly Padda**, Er.Jaspreet Kaur***

*(Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering & Technology)
** (Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering & Technology)
*** (Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering & Technology)

Abstract

Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data. Within these masses of data lies hidden information of strategic importance. The newest answer is data mining. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. As emphasized in the section on THE DATA MINING PROCESS, collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. We must build a predictive model based on patterns determined from known results, and then test that model on results outside the original sample.

Keywords: **Data mining, Data Process Model, Predictive Data Mining**

I. Introduction

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

II. Predictive Data Mining

The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on data collected from a variety of sources which may include corporate transactions, customer histories and demographic information, process control data, and relevant external databases such as credit bureau information or weather data. The result of the model building is a description of patterns and relationships in the data that can be confidently used for prediction.

To avoid confusing the different aspects of data mining, it helps to envision a hierarchy of the choices and decisions we need to make before we start:

- Business goal
- Type of prediction
- Model type
- Algorithm
- Product

At the highest level is the **business goal:** what is the ultimate purpose of mining this data? For example, seeking patterns in our data to help we retain good customers; we might build one model to predict customer profitability and a second model to identify customers likely to leave.

The next step is deciding on the **type of prediction** that's most appropriate: (1) *classification*: predicting into what category or class a case falls, or (2) *regression*: predicting what number value a variable will have (if it's a variable that varies with time, it's called *time series* prediction).

Classification

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave.

Regression

Regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques such as linear regression. Unfortunately, many real-world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification And Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural nets too can create both classification and regression models.

Model type: A neural net to perform the regression, perhaps, and a decision tree for the classification. There are also traditional statistical models to choose from such as logistic regression, discriminant analysis, or general linear models. Many **algorithms** are available to build models. When selecting a data mining **product,** be aware that they generally have different implementations of a particular algorithm even when they identify it with the same name. These implementation differences can affect operational characteristics such as memory usage and data storage, as well as performance characteristics such as speed and accuracy.

III. Data mining Process

Process Models

Recognizing that a systematic approach is essential to successful data mining. A process model designed to guide the user through a sequence of steps that will lead to good results.

The Two Crows Process Model

The data mining process is not linear — we will inevitably need to loop back to previous steps. For example, what we learn in the "explore data" step may require we to add new data to the data mining database. The initial models we build may provide insights that lead we to create new variables.

The basic steps of data mining for knowledge discovery are:

1. Define business problem
2. Build data mining database
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model and results

Let's go through these steps to better understand the knowledge discovery process.

1. Define the business problem. First and foremost, the prerequisite to knowledge discovery is understanding data and business. Without this understanding, no algorithm, regardless of sophistication, is going to provide us with a result in which we should have confidence. Without this background we will not be able to identify the problems we're trying to solve, prepare the data for mining, or correctly interpret the results.

2. Build a data mining database. This step along with the next two constitute the core of the data preparation. Together, they take more time and effort than all the other steps combined. There may be repeated iterations of the data preparation. *These data preparation steps may take anywhere from 50% to 90% of the time and effort of the entire knowledge discovery process!*

The data to be mined should be collected in a database. Depending on the amount of the data, the complexity of the data, and the uses to which it is to be put, a flat file or even a spreadsheet may be adequate. To store this data in a different DBMS with a different physical design than the one we use for our corporate data warehouse. Increasingly, people are selecting special purpose DBMSs which support these data mining requirements quite well.

The tasks in building a data mining database are:

a. Data collection
b. Data description
c. Selection
d. Data quality assessment
e. Consolidation and integration
f. Metadata construction
g. Load the data mining database
h. Maintain the data mining database

a. Data collection. Identify the sources of the data we will be mining. A data-gathering phase may be necessary because some of the data we need may never have been collected. We may need to acquire external data from public databases or proprietary databases .A Data Collection Report lists the properties of the different source data sets. Some of the elements in this report should include:

- •Person/organization responsible for maintaining data.
- • DBA
- •Storage organization (e.g., Oracle database, VSAM file, etc.)
- • Size in tables, rows, records, etc.
- • Size in bytes
- • Security requirements
- • Restrictions on use

- Privacy requirements

b. Data Description Describe the contents of each file or database table. Some of the properties documented in a Data Description Report are:

- Number of fields/columns
- Field names

For each field:

- Data type
- Definition
- Description
- Source of field
- Unit of measure
- Specific time data (e.g., every Monday or every Tuesday)
- Primary key/foreign key relationships

c. Selection. The next step in preparing the data mining database is to select the subset of data to mine. This is *not* the same as sampling the database or choosing predictor variables. Rather, it is a gross elimination of irrelevant or unneeded data. Other criteria for excluding data may include resource constraints, cost, restrictions on data use, or quality problems.

d. Data quality assessment. GIGO (Garbage In, Garbage Out) is quite applicable to data mining, so if we want good models we need to have good data. A data quality assessment identifies characteristics of the data that will affect the model quality. There are a number of types of data quality problems. Single fields may have an incorrect value. Inconsistencies must be identified and removed when consolidating data from *multiple sources.*

e. Integration and consolidation. The data we need may reside in a single database or in multiple databases. The source databases may be transaction databases used by the operational systems of our company. Other data may be in data warehouses or data marts built for specific purposes. Data integration and consolidation combines data from different sources into a single mining database and requires reconciling differences in data values from the various sources. Improperly reconciled data is a major source of quality problems.

f. Metadata construction. The information in the Dataset Description and Data Description reports is the basis for the metadata infrastructure. In essence, this is a database about the database itself. It provides information that will be used in the creation of the physical database as well as information that will be used by analysts in understanding the data and building the models.

g. Load the data mining database. In most cases the data should be stored in its own database. For large amounts or complex data, this will usually be a DBMS as opposed to a flat file. Having collected, integrated and cleaned the data, it is now necessary to actually load the database itself. Depending on the DBMS and hardware being used, the amount of data, and the complexity of the database design, this may turn out to be a serious undertaking that requires the expertise of information systems professionals.

h. Maintain the data mining database. Once created, a database needs to be cared for. It needs to be backed up periodically; its performance should be monitored; and it may need occasional reorganization to reclaim disk storage or to improve performance. For a large, complex database stored in a DBMS, the maintenance may also require the services of information systems professionals.

3. Explore the data. The goal is to identify the most important fields in predicting an outcome, and determine which derived values may be useful. In a data set with hundreds or even thousands of columns, exploring the data can be as time consuming and labor-intensive as it is illuminating. A good interface and fast computer response are very important in this phase because the very nature of we exploration is changed when we have to wait even 20 minutes for some graphs.

4. Prepare data for modeling. This is the final data preparation step before building models. There are four main parts to this step:

a. Select variables
b. Select rows
c. Construct new variables
d. Transform variables

a. Select variables. Ideally, we would take all the variables we have, feed them to the data mining tool and let it find those which are the best predictors. In practice, this doesn't work very well. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models. A very common error, for example, is to use as a predictor variable data that can only be known if we know the value of the response variable. While in principle some data mining algorithms will automatically ignore irrelevant variables and properly account for related columns.

b. Select rows. As in the case of selecting variables, we would like to use all the rows we have to build models. Consequently it is often a good idea to *sample* the data when the database is large. This yields no loss of information for most business problems, although sample selection must be done

carefully to ensure the sample is truly random.

c. Construct new variables. It is often necessary to construct new predictors derived from the raw data. Certain variables that have little effect alone may need to be combined with others, using various arithmetic or algebraic operations.

d. Transform variables. The tool we choose may dictate how we represent our data, for instance, the categorical explosion required by neural nets.

5. Data mining model building. The most important thing to remember about model building is that it is an iterative process. We will need to explore alternative models to find the one that is most useful in solving our business problem. The process of building predictive models requires a well-defined training and validation protocol in order to insure the most accurate and robust predictions. This kind of protocol is sometimes called supervised learning.

6. Evaluation and interpretation. It includes the following:

a. Model Validation. After building a model, we must evaluate its results and interpret their significance. More importantly, accuracy by itself is not necessarily the right metric for selecting the best model. We need to know more about the type of errors and the costs associated with them

b. External validation. As pointed out above, no matter how good the accuracy of a model is estimated to be, there is no guarantee that it reflects the real world. A valid model is not necessarily a correct model. One of the main reasons for this problem is that there are always assumptions implicit in the model.

7. Deploy the model and results. Once a data mining model is built and validated, it can be used in one of two main ways. The first way is for an analyst to recommend actions based on simply viewing the model and its results. For example, the analyst may look at the clusters the model has identified, the rules that define the model that depict the effect of the model.

The second way is to apply the model to different data sets. The model could be used to flag records based on their classification, or assign a score such as the probability of an action. Or the model can select some records from the database and subject these to further analyses with an OLAP tool.

IV. Conclusion:

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of Applications, from improving revenues to reducing costs. Building models is only one step in knowledge discovery. It's vital to properly collect and prepare the data, and to check the models against the real world. The "best" model is often found after building models of several different types, or by trying different technologies or algorithms. Choosing the right data mining products means finding a tool with good basic capabilities, an interface that matches the skill level of the people who'll be using it, and features relevant to our specific business problems. After we've narrowed down the list of potential solutions, get a hands-on trial of the likeliest ones.

References:

[1] DM Review, May 2003 "Data Mining in Depth" column: "Using Data Mining to Find Terrorists,"

[2] "Pan for Gold in the Clickstream," Information Week, March 12, 2001

[3] White paper: "Building Profitable Customer Relationships with Data Mining," in PDF format (156 KB)

[4] "Mining Large Databases -- A Case Study" (written for IBM), in PDF format (250 KB)

[5] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining".

[6] John Wiley & Sons *Data Mining: Concepts, Models, Methods, and Algorithms*. Kantardzic, Mehmed (2003).

[7] Alex Guazzelli, Wen-Ching Lin, Tridivesh Jena. PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics. CreateSpace, 2010

[8] Berry, Jonathan Database Marketing *Business Week* (September 5, 1994):56-62

[9] Cipolla, Emil T. ,Data Mining: Techniques to Gain Insight Into Your Data *Enterprise Systems Journal* (December 1995):18-24, 64

[10] Conner,Louis Mining for Data *Communications Week* (February 12, 1996):37-41