WHITE PAPER: UNDERSTANDING DATA MINING

by Michael J. Cavaretta, PhD

What is Data Mining?

Data Mining is a process for discovering data relationships hidden in large databases. Many companies have stored critical business data over a number of years and attempted to use it to aid in decision making. It is easy to get facts out of a database. For example, at this day and time, a customer purchased item X from store 123. What's desired is knowledge. For example, stores 123, and 130, sell 30% more of item X than all other stores. Generally speaking, the more specific the knowledge, the more valuable it is for decision making. Data Mining is a process for the discovery of this valuable business knowledge.

Why use Data Mining?

Data Mining can help businesses better understand how they work. For example, one technique is market basket analysis. This is concerned with finding customer preferences to aid in increased customer satisfaction and profitability. The Data Mining process is used to determine customer buying patterns implicit in the data. For instance, customers who buy item X are likely to buy item Y at the same time. Possible decisions based on this knowledge could be to place these items close together in the store or running a promotion on one of these items to increase the sales of both items.

Where has Data Mining been applied?

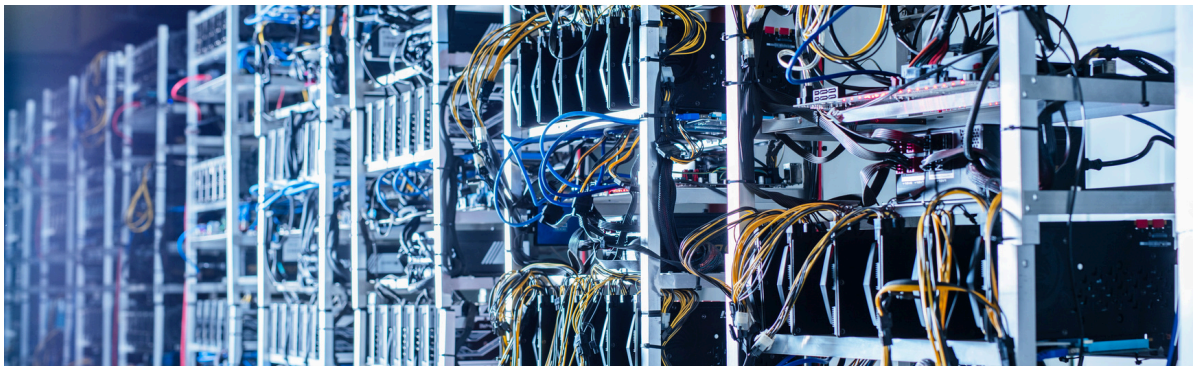
Data Mining applications have been fielded in a number of areas. A sample of these industries and their corresponding applications include:

- Retail: determining items for cross-promotions, store site selection, market basket analysis, forecasting demand
- Marketing: finding market segments, finding customer buying trends
- Finance: discovering expert system rules for underwriting, classifying accounts receivable by collection potential, forecasting price changes in foreign currency markets
- Healthcare: determining patient outcomes, analyzing managed care contracts
- Manufacturing: component failure diagnosis

What technologies are used in Data Mining?

Most Data Mining tools are structured around two techniques; machine learning and visualization. Visualization is the visual representation of data. The power of visualization is its ability to graphically represent data values. Changing the graphical representation of the data by changing colors, shapes, or other graphical elements, makes data relationships more easily discerned. The power of machine learning techniques lies in their ability to examine many more data relationships than a person could. The use of visualization and machine learning techniques are complementary in Data Mining. Visualization is generally used to look for outliers, overall trends and relationships and aid in data extraction at the beginning of the project. Machine learning is used later in the project to find relationships when the project has become more focused.

In Data Mining, the three most popular machine learning representations are decision trees, production rules, and neural networks. Decision trees classify the data by using data factors to successively breakdown the data elements into smaller and smaller groups. Production rules classify the data by using a set of expert-system-like rules. Production rules can be generated by using a search process to try combinations of rules or by extracting the rules from decision trees. In neural networks, knowledge is represented as links connecting a set of nodes. The strength of these links indicates the relationships between the data factors.



There are advantages and disadvantages to each of these representations. The advantage of decision tree and production rule representations is that they can be read like English sentences. However, with a large number of data factors it might be very difficult to actually understand what is being said. The disadvantage of these representations is that they are not suited for numbers covering large intervals. This is because each rule or point in the decision tree represents one relationship. To represent the relationships over a large interval many rules or decision tree points are required. The advantage of neural networks is that they can represent a numeric relationship over large intervals with a compact representation. Their disadvantage is that the representation cannot be read like production rules or decision trees.

What kind of Data Mining tools are there?

There are a large number of tools to help with Data Mining projects. These tools range from publicly available visualization and machine learning algorithms to highly complex packages, using multiple machine learning and visualization strategies running on parallel processing machines costing hundreds of thousands of dollars. Finding the best tool(s) for a Data Mining project is dependent upon a number of conditions, such as, the purpose of the project (e.g., market basket analysis), and the size of the database to be mined. Flexibility is very important in choosing Data Mining tools and algorithms, as using different strategies can give different perspectives on data relationships.

What is required to build a Data Mining application?

There are a number of steps required in building a Data Mining application.

1. Determine the scope of the Data Mining project. The scope of the project determines the data elements that must be collected for a successful project. It is important to focus the project on an objective business purpose.
2. Build the Data Mining database. The data required to answer the questions in the previous step may be scattered in many different databases, or not stored on a computer at all. Data stored in different databases may need to be consolidated and any discrepancies must be examined. Additionally, the data may need to be "cleaned" to eliminate errors. Steps one and two can require 50% or more of the total time spent on the project.
3. Quantify the data elements. Is a "large spender" defined as $50 per week or $300 per month? Is grouping clothes washers and ovens together more meaningful than separating them? Close association with domain experts will help specify the data elements that are meaningful from a business orientation.
4. Use Data Mining algorithms to determine relationships in the data. A number of different algorithms may be required to find the desired relationships. Some algorithms may be appropriate in the early stages of Data Mining, others in the later stages. In certain cases, using several Data Mining algorithms in parallel will provide valuable insight from different viewpoints.
5. Analyze the relationships found in the previous step for their applicability to the scope of the project. This step will likely require a domain expert. The domain expert will indicate whether the relationships are too specific or too general and suggest areas worthy of further investigation.
6. Present the results. A report specifying meaningful relationships is useful. However, a report provides only a one-time benefit. An application that allows domain experts to be creative in discovering relationships is even more useful. If a Data Mining application is being delivered, training on the application as well as how to find relationships in the data is required.

The concentration of the initial Data Mining prototype is on reducing errors in the database (i.e., steps one, two, three and five). A number of iterations may be required to gain understanding in the particulars of the data to be mined. In later prototypes the concentration will switch to steps three, four, and five. Other factors affecting the breakdown of time required for a Data Mining project include, the final application and the existence and condition of the data warehouse. For example, a sales forecasting application is used for prediction. Once the data relationships are found they can be used until the business changes. Conversely, in market basket analysis a company is constantly looking for new data relationships. Over a complete sales forecasting project more time is spent in steps one through three while in market basket analysis more time is spent in steps four through six.

What is the conclusion?

Many companies are looking to explore the data generated from their everyday transactions. Meaningful relationships can be discovered using machine learning and visualization technologies. Data mining applications utilizing these techniques have been successful in a number of different fields including retail and marketing. These Data Mining applications allow companies to discover knowledge that can give them a competitive edge.