

ezDataMiner and the Strategic Advantages of Data Mining

Moheb A Kasem, Chris Bassell, Dean Amo, Andrew Jambor,
Marianne J. D'Onofrio, Olga Petkova
Central Connecticut State University, USA

“Most organizations can be currently labeled 'data rich', since they are collecting increasing volumes of data about business processes and resources. Typically, these data mountains are used to provide endless 'facts and figures' such as 'there are 60 categories of occupation', '2000 mortgage accounts are in arrears' etc. Such 'facts and figures' do not represent knowledge but if anything can lead to 'information overload'. However, patterns in the data represent knowledge and most organizations nowadays can be labeled 'knowledge poor'.”

--- Akeel Al-Attar, *Data Mining - Beyond Algorithms*

Abstract

The purpose of the project is to discuss the data mining concept and the strategic advantages that it presents to organizations of all sizes. The project also covers the various components used in the data mining environment such as data store, data warehouse and data mart. Furthermore, the project documentation educates new users about the data mining concept and the basic requirements for adopting such a technology. The documentation arms new users with the essential criteria of evaluating a new data mining tool that can be implemented to mine the business information. In addition, the documentation introduces, ezDataMiner, the team's own web data mining application. ezDataMiner is an open source data miner that was designed to utilize WEKA's line command data miner tool.

The documentation includes the web data mining software system manual including basic hardware and software requirements and a detailed implementation guide. Also, an interactive web tutorial for the system and the data mining process is part of the project to guide the first time user through the application functions.

Data mining techniques such as classification, market basket analysis, sequencing and clustering are explained with some applied examples. Several case studies are presented to demonstrate the benefits of applying data mining in a data focused business. Several variations of the Online Analytical Processing are analyzed to demonstrate how they relate to the process of data mining.

Finally, the documentation discusses the future of data mining and some of the controversial issues surrounding the data mining process such as data ownership, privacy, security, and individual privacy rights.

Data Mining Overview

The amount of operational data collected and stored by organizations is steadily increasing. Adriaans and Zantinge state that the amount of information stored as data in the

world doubles roughly every 20 months. This proliferation of information is literally causing organizations to drown in their own data. But at the same time, organizations and individuals continue to be starved for knowledge. The key is to seek hidden nuggets of valuable, strategic information in this data thus converting it into knowledge helping to curb this craving. Data mining and data analytic tools allow organizations to probe into the mass of collected data and effectively uncover these nuggets of new information which can aid strategic decisions as well as provide feedback critical to monitoring business performance. These fruits of data mining can effectively feed an organization's Decision Support System (DSS) supporting all aspects of the business (Adriaans & Zantinge, 1996). Data mining accomplishes this task by applying statistical algorithms to data which assist analysts in finding meaningful patterns and relationships. These patterns and relationships in turn can be used to both increase revenues and reduce costs for organizations (Berry & Linoff, 2000).

Successful results can be difficult to obtain from data mining technology, but when this goal is achieved it can lead the way to sustained competitive advantage for the organization. This success is hard to obtain because technology alone is not enough. The organization must become information oriented, integrating and acting upon discoveries to receive their full benefit. The key to successful data mining is to have an idea about the expected results before starting the mining process. These expected results can be expressed in a form of questions that need to be answered. For example, if a mortgage company is mining financial data, a good expectation to set is to find the common characteristics of customers who default on their loans. Setting an expectation helps evaluate the success of the mining results. Although data analytics can be successfully applied to just about any aggregate body of knowledge, the focus of this document is on data mining as it applies to businesses, e-commerce and management information systems (Two Crows Corporation Report, 1999).

Data Warehouses

The Two Crows Corporation report addresses the data warehouse as a huge factor in preparing and manipulating the data before applying the mining process. When discussing data quality, the term data warehouse is likely to surface as a result of the data preparation services that it provides. The creation of a data warehouse can require a major percentage of the resources for the data mining initiative.

Where does the data come from to load the data warehouse? Data comes mainly from daily operational transactions and business information systems that are stored in application databases. With the growth of the Internet, huge amounts of data can be generated through a company's web site that is integrated with a back-end database. To further complicate matters, data from these different types of systems are distributed throughout the business units of an organization and are built using a wide range of technologies (Two Crows Corporation Report, 1999).

The data warehouse is usually built when an organization is trying to manage two or more of its main data marts. The data store and the data mart are the building blocks for having a data warehouse.

The Data Store

The most common component of the data warehouse architecture is the operational data store (ODS). The ODS is populated by applications with information that is usually subject

oriented, changeable and more or less current. Some of the main focuses of the content include customers, transactions, products and assets. Each business unit is the owner of its operational data store. An ODS may not only be internal to the organization but may also be located externally. For instance, if a business was lacking certain information it needed to make key decisions, it could tie into and populate its ODS with an external ODS owned by another entity. This would in turn feed the organizations own data warehouse (Two Crows Corporation Report, 1999)

Data Marts

When discussing data warehouses, the term data mart needs to be introduced. The two terms are sometimes used interchangeably but the term data mart actually refers to a smaller scale, subject focused version of a data warehouse. Sometimes it can be cost prohibitive or too discouraging (at least at first) to aggregate an entire organization's data into a central data warehouse. A data mart could be a profitable, lower risk alternative way to proceed on a data mining initiative while still leaving the possibility of upgrading to a fully integrated data warehouse system over time. One of the many advantages to starting on a relatively small scale is that the organization can concentrate on creating the data mart with just one or more of the organization's business units. It is much easier to deal with one management group with their specific goals and their data than it is to fulfill all the needs of the entire organization. Furthermore, since data mining is a process, not an event, things that are learned and mistakes that are made along the way are on a relatively small and manageable scale (Two Crows Corporation Report, 1999).

Data Warehouse

The Two Crows Corporation report has a lengthy definition for data warehouse; but first we introduce this simple, yet concise, definition of the data warehouse concept. The data warehouse is an integrated view of an organization's dispersed operational data stores. There are several key attributes of a data warehouse that make it a suitable staging ground for data mining operations. First, each unit of data is read-only and is relevant to some point in time (Cabena et. al., 1998). Users must update information through the operational systems from which the data warehouse is loaded, helping to ensure there are no inconsistencies between the two systems. Furthermore, the data within the warehouse is separate from the data used in day-to-day operations. This is a key point when thinking about performance degradation of ODSs responsible for daily transactions, due to the high volume of processor time needed for data mining activities. Typically the data warehouse has a large volume of data stored in it as it contains historical data. The data within the data warehouse is organized around the major subject areas of the business such as inventory, sales and marketing. All operational data is not loaded; only data related to business decisions is kept in the warehouse. Major roles of the data warehouse include:

- Standardizing data types and sizes from diverse data stores
- Summarizing operational data
- Storing historical data

There are only two data operations ever performed in the data warehouse: data loading and data access. Before data is loaded from the ODSs, whether they are internal or external to the organization the data must be filtered and cleansed. This process allows only data needed to

support the Decision Support System into the data warehouse and insures that the data is consistent and integrated. For example, it would not be appropriate to have different measurement attributes associated with the same types of data in the data warehouse. An example would be monetary attributes. One business unit located in the UK would measure their monetary element's based on the Euro while another business unit in the U.S. would be based on the dollar. This data must be converted into a common unit of measure before loading into the data warehouse. Therefore, it is important that the data be conditioned for consistency before loading into the data warehouse. In addition, there may be several definitions of the same entity in operational databases; the data warehouse consolidates these into a single view of the entity. For example, divisions of an organization may service the same customer and each may have a different identifier for that customer. The data warehouse would assign a common number and map all divisional references to the consolidated view.

The most popular method for loading data warehouses is the extract, transform and load method (ETL). This is a frequent technique used by commercial data warehousing packages to load diverse data structures into a common format (Two Crows Corporation Report, 1999).

The Data Mining Process

ezDataMiner Process Model

The purposed data mining model is implemented in a software application that utilizes the WEKA algorithms with a user friendly web interface. WEKA is an open source data mining reference application with a wide range of mining algorithms implemented. The process model is explained in more details in the user reference manual, but below are the steps that a user is going to take in order to mine a file using the application.

1. Define business problem and expected results:
To make the best use of the data mining model, the user must make a clear statement of the objectives. In a hospital situation for example, the goal is to mine the patient file to find out if there are common factors that cause a certain diagnoses. Also, estimate the ratio of the number of patients per doctor in certain zip codes. In car sales situation, the goal is evaluate the customer attributes to identify an efficient target market.
2. Build data mining database (warehouse):
By copying all the files that need to be mined into a data mart first, it will be easy to perform the next two steps. It is good practice to perform the data mining in a separate data mart from the production database to avoid performance issues. The data portion (steps 2, 3, and 4) is the most time consuming of the entire process.
3. Explore data (run queries to summarize the stats about the database):
The goal is to identify the most important fields in predicting an outcome, and determine which derived values may be useful.
Depending on the size of data, exploring the data can be as time consuming and labor-intensive as it is illuminating. A good interface and a fast computer response are very important in this phase because the very nature of your exploration is changed when you have to wait even 20 minutes, let alone hours, for some graphs.
4. Prepare data for modeling:
This step involves converting the data to an acceptable format. The program will accept Comma Separated Value (CSV) files only. After making sure that the data is cleansed

and collected in one large file, the data should be saved in a CSV file format. Having “clean” data to mine is important for accuracy and desirable outcomes.

5. Mine the Data (execute the program):
Utilizing the Web User Interface, the user will be able to direct the program to the input file, set the file properties, choose the mining algorithm and trigger the mining process.
6. Evaluate the output (analyze the results):
To accurately analyze the results, one should first validate some of the outcomes against the queries and statistics that were collected before starting the process. Analyze the outcomes for useful conclusions to make use of the next step.
7. Deploy model and results (making appropriate decisions based on the outcomes)
Once a data mining model is built and validated, it can be used in one of two main ways. The first is for an analyst to recommend actions based on simply viewing the model and its results. The second way is to apply the model to different data sets.

Machine Learning

“Machine learning is the automation of a learning process where learning is tantamount to the construction of rules based on observations of environmental states and transitions. This is a broad field that includes not only learning from example, but also reinforcement learning, learning with teacher, etc. A learning algorithm takes the data set and its accompanying information as input and returns a statement; e.g. a concept representing the results of learning as output. Machine learning examines previous examples and their outcomes and learns how to reproduce these and make generalizations about new cases.

Generally a machine learning system does not use single observations of its environment but an entire finite set called the training set at once. This set contains examples of observations coded in some machine-readable form. The training set is finite hence not all concepts can be learned exactly” (Parallel Computer Center, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_2.html, 2004).

Data Mining Applications

Applications for Data Mining

Many organizations have taken advantage of the data mining concept and the available data mining tools to gain a competitive advantage. The list below shows the industries that are more likely to implement data mining activities. The common denominator in all of these industries is automation and data collection. Also, fierce competition in their respective fields is a major impetus driving the adoption of data mining tools and techniques (Parallel Computer Center, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_2.html, 2004).

Retail/Marketing

Data mining applications can be applied in the retail and marketing industries to accomplish very important tasks and provide excellent advantages. These advantages can be in the form of:

- Identifying buying patterns of customers.
- Finding associations among customer demographic characteristics.

- Identifying target markets.
- Predicting response to mailing campaigns.
- Performing market basket analysis.

Banking and Other Financial Institutions

The financial institutions, such as banks and investment companies, are the pioneers in taking advantage of data mining. They were able to use it to do things like:

- Detecting patterns of fraudulent credit card use.
- Identifying 'loyal' customers.
- Predicting customers likely to change their credit card affiliation.
- Determining credit card spending by customer groups.
- Finding hidden correlations between different financial indicators.
- Identifying stock trading rules from historical market data.

Insurance and Health Care

Detecting insurance fraud used to cost insurance carriers a lot of money in the past. Now the insurance companies can mine client's data to detect certain suspicious habits. For example they used data mining in:

- Conducting claims analysis - i.e. which medical procedures are claimed together.
- Predicting which customers will buy new policies.
- Identifying behavior patterns of risky customers.
- Identifying fraudulent behavior.

Transportation

The transportation industry was able to use data mining to determine distribution schedules among outlets and analyze loading patterns. This initiative allowed the insurance industry to minimize the amount of travel by assigning distribution centers to more centralized vendors.

Medical Health Care

One of the biggest applications for the data mining tools is the medical healthcare industry. It used to take researchers many hours to collect information into books and matrixes to try to figure out the cause of disease. With data mining the medical healthcare industry was able to:

- Characterize patient behavior to predict office visits.

- Identify successful medical therapies for different illnesses.
- Analyze cause and effect of diseases.

Telecommunications

The telecommunication industry was able to detect fraudulent use of services and locate lost or stolen products by applying data mining techniques.

Credit cards

Also, in the credit card industry; the providers were able to detect fraudulent credit card use early in the process of misuse. Also credit card providers were able to model the shopping patterns of customers and use it for efficient advertising. Actually, the collected data became marketable and the credit card companies were able generate revenue from selling the information to third parties.

The Strategic Value of Data Mining

Data mining can be used to improve customer relationship management (CRM). Paul Duke explains that often the goals of customer relationship management systems read like the holy grail of marketing. Some CRM application examples include:

- Identifying the most profitable new customers for direct mail marketing
- Identifying the most profitable existing customers for cross selling
- Reducing churn of existing customer-base

With the prevalence of the Internet, eCRM systems are integrating click-stream data with historical sales and demographic data to address real-time, online marketing issues such as personalized content, intelligent cross selling, price management and banner ad management. At the heart of these applications are a variety of data sources, often linked geographically. The underlying intelligence comes from predictive models that can accurately forecast future customer response based on historical customer response patterns (Duke, 2001).

Data mining can also be used to detect fraud. It can help a sports team to select the most productive player, or group of players, that simple statistics do not reflect. It can help a company decide what the best advertising strategy for its products should be. Data mining provides a model where a plan of action can be formulated and implemented. The results of the actions can then be measured and fed back into the data warehouse for future use. A company can use a well-crafted data mining plan of action and maintenance to continuously probe for advantages in the market place. The cost may very well be worth the return on investment. Justifying a data mining initiative is a difficult task because results are difficult to predict. Funds can easily be redirected to projects with more tangible benefits. A data mining project is high risk but has the potential to produce competitive advantage resulting in high reward. The business case below will outline justifications for a data mining initiative.

Adoption Issues

The major adoption issue with data mining initiatives is that the technology is still maturing and it carries a high-risk level. In addition, the organizational culture itself has to be ready to embrace the technology and act on the results. If the organization is not information oriented there is little chance for data mining success. There are several significant issues in implementing, using, and maintaining a data-mining capability that organizations need to address.

- **Understand what you are getting** - Implementing a data mining software engine requires a skilled staff and an IT infrastructure put in place.
- **Data readiness for analysis** - In order to maximize the benefits of the data mining analysis, the data being mined has to be as normalized as possible. Seventy to 85 percent of the work in building models using data mining relates to the cleaning and preparation of data prior to a specific analysis.
- **Staffing** - Data mining does not replace skilled employees. Also, data mining is not a one shot process; it is an incremental learning process that increases the quality of the output the more often it is used.
- **Taking action** - Organizations have to apply the recommendations of the data mining process in order to realize the return on investment. Many organizations forgo the recommendations because it sometimes requires change.
- **Financial considerations** - Just like a lot of the new IT systems, businesses do not appreciate the actual cost of data mining strategy. "The costs of integrating the data mining package and making it work in a specific environment are often underestimated by 50 percent or more" (Labovitz, 2003).

Applied Business Examples

This section will present areas in business where data mining is already being applied. This will serve as a starting point when attempting to apply data mining technology in an organization. One of the major applications for data mining is analyzing customer behaviors and predicting customer-spending trends.

The National Basketball Association (NBA)

Games and sports in particular are a good source for data. All kinds of statistics can be collected such as runs batted in (RBI's), earned run average (ERA), number of saves on goal, total yards gained on pass plays and so on. The National Basketball Association (NBA) is an abundant producer of statistics. During each of the 1200 or so games played throughout the season, members of the NBA's Game Stats program enter game statistics into a laptop computer. They keep busy during each game, which averages 200 changes of possession, with numerous plays and player offensive/defensive match-ups. Each piece of data that is entered gets a universal time stamp based on the game play clock. All this data is uploaded to an IBM Hosting Center that is running an IBM application called DB2 Universal Database. About 25 of the teams are using an IBM product called Advanced Scout. They tie into the database at the IBM Hosting Center and download the data to integrate into their own databases. Using Advanced Scout, they

can mine the data for meaningful patterns and relationships about the play in general and the players specifically.

In the 1997 NBA Finals, the number four-seed Orlando Magic was down two games to none in a best of seven series against the number two-seed Miami Heat. The coaches, using Advanced Scout data mining, found that certain combinations of players worked better together than others. They scored more points together and played better defense together. The coaches gave more playing time to these players, which turned the series around. They almost pulled off a historical NBA upset. Data mining gave the team a competitive edge that helped rally its fans behind them. This translated into increased season ticket sales for the next year. Another benefit for the coaches was a drastic decrease in the time spent going over game tapes. Coaches would typically spend weeks studying the tapes, but now when they spot interesting patterns or events with Advanced Scout, they can cue the universal time stamp on the data with a video of the game to get to the precise spot and analyze the play. Tom Sterner, Assistant Coach of the Orlando Magic says, "By helping us make better decisions, Advanced Scout is playing a huge role in establishing incredible fan support and loyalty; that means millions of dollars in gate traffic, television sales and licensing" (IBM web site, <http://www-1.ibm.com/industries/media/doc/content/casestudy/356304111.html>, 2004).

Eddie Bauer

Eddie Bauer, a multi-channel apparel retailer, is using data mining products from SAS to build better one-to-one customer relationships with its more than 15 million retail, catalog and Internet customers. Customer Relationship Management is a top priority for the firm. In the past, the company used a channel-centric viewpoint to guide company direction and make decisions. SAS helped Eddie Bauer to see that it would be more profitable to use a customer-centric view to feed information into its Decision Support System. By using this approach, the company now analyzes data from all channels to better understand their customer behavior. One of the problems the company faced prior to SAS was that their data was scattered throughout different ODSs. SAS was charged with integrating the data into a data warehouse thereby providing more efficient and effective data mining and analysis. Eddie Bauer uses SAS to perform predictive modeling to determine things like customer loyalty, seasonal buying habits, which customers are most likely to buy and targeted advertising. They use data mining to make tracking and maintaining the customer base easier and more efficient (SAS/COM Magazine, <http://www.sas.com/success/eddiebauer.html> 2001).

Carrier Corporation

Carrier Corporation, a part of the United Technologies Group, headquartered in Farmington, CT, uses data mining technology to profile its online customers in an effort to provide them with direct and focused marketing promotions. Carrier's goals were to increase awareness for the web store and increase traffic and sales on the site. They turned to WebMiner Inc., a data mining technology provider, to help boost revenue up from \$1.47 per visitor. WebMiner took a years worth of Carrier's online data along with a database of users who had signed up for an online sweepstakes and combined it with a demographics database. The technology team mined the information using classification and association techniques to come up with a profile of Carriers' online customers. This was information that Carrier had no idea about prior to the data mining initiative; they barely knew anything about their online customers.

The profiles were matched with zip codes to produce a predictive model so that new visitors to Carrier's web store only have to type in their zip code to get a customized pop up window that offers appropriate products. Officials at Carrier say that that was the first time they have delivered intelligent data-driven promotions. They found that money spent on radio ads would be better spent on subway advertising based on the typical online customer profile. The plans put into action based on the results of data mining increased the revenue per visitor to \$37.42. The cost to the company for WebMiner services was a one time \$10,000 installation fee and a \$5 fee for each unit sold on the web site to support the continuing services of WebMiner (Whiting, 2001).

Data Mining Market Analysis

Many commercial data mining offerings are available from established and emerging vendors for nominal prices. Some are actually offered for free. IBM – Intelligent Miner cost for licenses range from \$25,000 to \$150,000 U.S. depending on the version and scale of information being mined (IBM online software catalog, 2004).

Other vendors like Thinking Machines, DataMind, Business Objects (formally Crystal Reports), SAS Institute, and Cognos have their own data mining tools. Some of these tools are sold as stand alone applications and others are packaged with the main vendor product.

Open Source Initiatives

WEKA

As mentioned earlier WEKA is an open source collection of machine learning algorithms for data mining with a wide range of these mining algorithms implemented. WEKA's data mining implementation is released under the GNU General Public License in order to maximize the end goal of interoperability and widen its acceptance. Within the WEKA package there are tools for cleaning data, mining data and visualization of results. The application contained within this project integrates only the open source data mining algorithms in order to take advantage of proven mining technology. WEKA stands for Waikato Environment for Knowledge Analysis and was developed at the University of Waikato in New Zealand.

Other Open Source Initiatives

Standards are just beginning to emerge for data mining technology and larger software vendors are beginning to join these standards groups. There was previously little to no support for standards in data mining from larger software vendors backing these standardization initiatives. Names like Oracle, Sun Microsystems and IBM Corporation are just a few of these large software vendors that are beginning to organize and standardize the way data mining applications integrate. Other open source initiatives include:

- PMML – a XML formatted standard for data mining output to enable any PMML compliant visualization tool to render results. PMML was created by the Data Mining Group, DMG.org, whose members include names such as IBM Corporation, Microsoft, Oracle and many others.
- JSR-73 is an open source data mining algorithm application programming interface (API) which includes a sample implementation. This allows developers to implement or modify algorithms while allowing these algorithms to be compatible with any user interface that is JSR-73 compliant.

ezDataMiner Overview

The purpose of this chapter is to introduce the new user to the world of data mining with a focus on using the ezDataMiner application. It is not meant to be an exhaustive discussion on data mining, but merely an introduction to get one started in using ezDataMiner. After reading this document, you should have a good understanding of some of the basic terms used in data mining and how to best prepare your data for the mining process.

One of the most important aspects of data mining is to know your data. The first step in getting to know your data is to be familiar with the data mining terminology used.

Terminology

Two important terms you will see throughout are ‘attribute’ and ‘value’. These two terms are related. You can think of an attribute as the name of a column heading in your data set. Take for example the weather data set included with the application and shown in table 1.

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	FALSE	No
Sunny	80	90	TRUE	No
Overcast	83	86	FALSE	Yes
Rainy	70	96	FALSE	Yes
Rainy	68	80	FALSE	Yes
Rainy	65	70	TRUE	No
Overcast	64	65	TRUE	Yes
Sunny	72	95	FALSE	No
Sunny	69	70	FALSE	Yes
Rainy	75	80	FALSE	Yes
Sunny	75	70	TRUE	yes
Overcast	72	90	TRUE	Yes
Overcast	81	75	FALSE	Yes
Rainy	71	91	TRUE	No

Table 1: The weather file

The very first row or header of each column is an attribute name; outlook, temperature, humidity, windy, and play are all attributes of the weather data set, therefore this data set has five attributes. The data in each of these columns are the values of their respective attribute. For instance, the outlook attribute has three distinct values: sunny, rainy and overcast. When an attribute has a distinctly named set of limited values, they are called nominal values. The nominal attribute windy has two values which are given as true and false. Both the temperature and humidity attributes have integers that represent a value that can vary. These are called numeric values. Numeric attribute values are treated differently than nominal values in data mining. It is common for data mining algorithms to use various mathematical operations such as regression analysis, less than/greater than calculations, standard deviation etc. on values of attributes that are declared as numeric. Attribute types are discussed in more detail later in this text.

Another important term or concept is that of ‘class value’. In the context of ezDataMiner, the class value is the concept to be learned. It is used in all of the classification algorithms in ezDataMiner. For the weather data, the concept to be learned is whether or not to play (i.e. play = yes or play = no) given the weather conditions. So in this case, the play attribute is called the class value. The class value is chosen in ezDataMiner by the user after they have selected which

attributes to include in the analysis. There can be only one class value per data set in ezDataMiner. The class value is ignored when using association algorithms such as Aprori and Prism in ezDataMiner.

Another term frequently used in data mining is the word 'instance'. This is another name for a row of data or a tuple or a record in the data set. Each instance in a data set is characterized by the values of attributes that measure different aspects of the instance. One can think of each instance as an individual, independent example of the concept to be learned. There are fourteen instances of information in the weather data set shown above.

What do you want to know?

Now that we have discussed the terminology, the next main consideration is to have an idea of what you want to know about the data. This is similar to the theme to your data mining endeavor. One has to ask the question: based on the data, what is it you want to know? It is not enough to just have a pile of data, throw it all together, and expect to run it through the various algorithms in ezDataMiner and come out with meaningful results. One has to have a game plan, so to speak. Whether it is trying to find associations of purchasing patterns in a data set of supermarket transactions (commonly known as market basket analysis) or trying to uncover certain rules for approving a bank loan using a historical database of loan approval information, you have to have a clue as to what you are after. Although this does not mean that you can not take a set of data that you know is related, or has a possibility of being related, put it all together and run it through ezDataMiner. The key here is that you know something about the data. In this case you know that there are, or possibly are, relations in the data at hand. The more you know about the data, the better the chances of choosing the right combinations of attributes that lead to good results.

Attribute Selection

As an extreme example of knowing the data, let us discuss attribute selection. Attribute selection is the process of including/removing certain attributes or columns of data for analysis. In ezDataMiner, there is a screen where one has the option of removing columns of data from the analysis. For example, you have been handed a data set from the company payroll database. Along with attributes such as salary, commission, bonuses and sales, there is an attribute called employee identification (a unique value assigned to each employee). These are pretty straight forward attributes to understand and without putting too much thought into it, one might think there could possibly be some relation between pay and sales (notice the first thing done here is to understand the data you are working with).

- Pop quiz: which attribute here would be a good choice for the class value?
(Answer is shown below)

In this case, the key is to realize that the employee identification attribute has nothing to do with the analysis. In fact, including the employee identification in the analysis will be extremely detrimental to the data mining effort. Sometimes, it is what is left out that has the greatest effect and best results. Again, it can not be stressed enough that you must know your data to increase the chances of success in any data mining operation. (Answer to pop quiz: if you said the class attribute should be sales, you're off to a good start).

Data Cleansing

One of the most important things to do, before starting the data mining process, is to make sure your data is as clean as possible. This is probably the most tedious, unglamorous and time consuming part of data mining. Experts say that up to 60% of a data mining effort can be spent cleansing data. But in order to get the best results possible, the data must be in good form, so to speak. Data cleansing involves making sure the values of attributes are consistent and normalized.

Data Consistency

Consistency of the data may involve several different aspects such as attribute value names being spelled the same throughout the data set or the format of a date value. Most data mining products don't know that variations on a name or date mean the same thing. For example, if you have a data set with an attribute listing the names of Connecticut towns, each must be spelled the same. A data mining product will consider West Hartford, W. Htfd. and W. Hartford as different values of the same attribute even though each means the same town. Also, nominal attribute values such as these are case sensitive. Dates or timestamps are another such value that can be input in different ways. More is discussed about dates and timestamps below in the ARFF Format section.

Data Normalcy

Normalizing data is important for coherent results. It is kind of like making sure one is comparing apples to apples as opposed to apples to oranges, so to speak. For instance, say one has a data set with international information in it to be mined. Also, let us say that there is an attribute representing the sales of different divisions in the data set. Through the process of getting to know the data, we might realize that the sales figures from the different divisions are stated in European terms for the UK divisions and American terms for North America. It would not be a good idea to leave the data as is, this is like comparing apples to oranges; the data needs to be normalized for an accurate analysis. This might involve converting the euros to dollars or vice versa. These kinds of problems are common for attribute values that measure quantity. The important thing to remember is that values of each attribute are represented in like terms.

Missing Values

It is common for data sets to have missing values. This happens for a number of reasons and could be significant unto it self. A missing value here and there might not mean much but if large amounts of data are missing, this could indicate a problem that needs to be investigated. Examining why so many values are missing can lead to revelations in and of it self. To say the least, the data mining effort will be compromised if too many values are missing. Even a few missing values in a small data set can have a detrimental effect on the results. That is to say, given a fixed number of missing values, it is better to have more data than less to work with. A nice feature of the ezDataMiner application is that any missing values in your data set are automatically formatted properly so that the mining algorithms can deal with them. So, that is one thing you don't have to worry about during the cleansing process. But, it is something to keep in mind and be aware of during the data mining process.

Outliers

Outliers are data that are out of the normal range relative to the rest of the data. This mostly concerns numeric data rather than nominal values. For example, using the weather data shown previously, if we examine the values for the temperature attribute, we observe that all of these values are within an acceptable range. But imagine if we ran across a value of 150. This temperature value would be considered an outlier in this case. There could be numerous reasons why this value shows up in the data set. The recording instrument used to measure the values could have malfunctioned, the person entering the data could have made a typo; the list goes on and on. The important thing to realize is that outliers, if used in the data set, can skew the results of the data mining application. This is where using graphing techniques such as scatter plots and histograms on ones data prior to the data mining process could help one identify and eliminate outliers.

The old adage of ‘garbage in – garbage out’ applies quite well to data mining. Strange and less than satisfying results happen when the data is not cleansed beforehand. Your chances for success diminish the less you know about your data and the dirtier your data is, therefore it is time well spent getting to know your data and going through and looking for inconsistencies.

ARFF Format

Attribute Relation File Format (ARFF) is the file format used in the ezDataMiner application. It is the layout that the data takes in order to be understood by the mining algorithms. Basically, what happens is that the comma separated value (CSV) file that is given to the application for data mining is transformed into an ARFF format. This is done automatically as one progress through the ezDataMiner application. The only input the user has in deciding the final outcome for the format is selecting attributes for analysis, choosing a class value, naming the relation and selecting attribute types. The two items that haven’t been discussed yet are naming the relation and selecting attribute types.

Naming the relation is straightforward. One can choose to enter any name they like. The relation name is used internally in the application and appears on the reports for identification purposes. It is not used in any calculations nor does it have any significance or plays a factor in the data mining results. The only thing to remember is that if the name one enters contains spaces, the whole string must be double quoted.

When one gets to the attribute type selection screen in the ezDataMiner application (not to be confused with the attribute selection screen where one selects which columns of data are to be included in the analysis) it is time to declare how one wants the various algorithms to handle the data. That is to say, it is time to describe the data in terms the algorithms understand.

The two basic data types are ‘numeric’ and ‘nominal’. There are also two other types available which are ‘date’ and ‘string’. Initially, the data types for all attributes are selected as nominal by default.

Nominal

The application determines automatically, based on up to the first 200 rows of each column of data in the data set, all unique values and lists these as the nominal values for each respective attribute. This is fine if the data contains only nominal attributes In that case, the user

doesn't have to determine the attribute types; just accept what is given and carry on. But, chances are that the data set contains numeric attributes also. In this case, there could be potentially 200 unique values listed across the attribute type selection screen. This could be disconcerting to some as they might be tempted to horizontally scroll to view all of the values. This is not the purpose of attribute type selection. The purpose is to decide what format they want to use for the attribute in question. All the user has to do is select/click anywhere on the row in question and a drop-down list will appear from which the user may select from the different data types.

Numeric

Numeric attribute values can be real or integer numbers. One thing to consider when dealing with attributes that have numeric values is that sometimes a number isn't a number; it is a label. Take for example a data set with labor figures. Besides having attributes such as production figures and processes one finds an attribute called 'shift'. The shift attribute may have the values 1, 2, or 3. One would not want to declare the shift attribute as numeric; it is a nominal value. The important thing to keep in mind about numeric attributes and their values is that data mining algorithms perform mathematical calculations on them. So, if you have an attribute with numbers, just ask yourself, would it make sense to analyze this attribute using math? If your answer is no, then leave it at its default nominal values.

Date

EzDataMiner requires that the date/time stamp fields be in this format "dd-MM-yyyy HH:mm:ss" (if there are spaces, such as between the date and time, you need to quote the entire thing as shown). The day, month and year can be switched around and different delimiters may be used (e.g. 12/31/2004) but be sure that both the data and attribute type declaration match exactly. For example, if the data set has a date attribute with values such as 12-31-2004, then the user should select the date attribute type followed by the date format used in the data. It should look like this: Date "mm/DD/yyyy" or if time is included with only hours and minutes: Date "mm/DD/yyyy HH:mm" The date attribute type is the only attribute where one has to specify a format after the type that matches the values in the data set.

String

String attributes allow one to create attribute values containing arbitrary textual values. This is very useful in text-mining applications, as one can create data sets with string attributes and values. Text mining is an emerging new area, and there are as yet no comprehensive sources of information. It is a complicated and involved process and above and beyond the scope of ezDataMiner. The string attribute type is included as a courtesy because the algorithms in ezDataMiner can handle this type of attribute, but no real explanation for its use (other than right here) is given. Although, if the reader is interested in working with strings or doing text mining, he/she is encouraged to do further research; the web is a good place to start.

What Algorithm for what Data?

The algorithm that is applied to a data set depends on what kind of data you have and what you want to know. Here are few general guidelines to get you started.

- If a class value is not readily apparent in the data set, then one of the association algorithms would be your best bet. Try both the Prism and Apriori algorithms.

- If you want to derive rules from the data set, such as if-then rules, then use the classification algorithms. OneR oftentimes comes up with surprisingly accurate and simple rules. Try the Decision Tree algorithm for a more comprehensive set of if-then rules.
- Another approach is to try all of the algorithms and see which one produces the best results based on the stratified cross-validation results (except for Apriori and Tertius which don't do stratified cross-validation, they use different factors to state the accuracy of their results).

Final Word

Data mining is a lot like looking for buried treasure. Sometimes you get lucky and hit the jackpot on the first try, but that is pretty rare. Most other times, one has to put in a lot of hard work and fool around with the data, by selecting and removing attributes to be analyzed or choosing a different class value, to get good results. And still other times, no matter how you slice and dice your data and what algorithm you use, you don't come up with anything useful. That is the way things are in data mining; there is no guarantee of success. But, when you do have success, the rewards can be substantial. Extremely useful knowledge can be gained through data mining endeavors that can't be realized using any other way. The most important thing is to understand and clean your data as well as possible and don't be afraid to experiment. So, now it is the time to go put on your mining helmet, light your carbide lantern and start mining. To obtain the rest of the documentation and the software files, please contact the authors directly.

References

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

Al-attar Akeel (2004, November, 19). *Data Mining - Beyond Algorithms*. Retrieved on November 19, 2004, from <http://www.attar.com/tutor/mining.htm>

Berry, M. J. A., & Linoff, G. S. (2000). *Mastering data mining, the Art and Science of Customer Relationship Management*. Wiley Computer Publishing, John Wiley & Sons, Inc.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining From Concept to Implementation*. Prentice-Hall.

Duke, P. (March/April, 2001). *Data Mining For Market Intelligence*, PC AI.

IBM, (2001, February 21). *NBA Coaches Score Big with IBM Data Mining Application*. Retrieved on September 8, 2004, from <http://www-1.ibm.com/industries/media/doc/content/casestudy/356304111.html>

Labovitz, Mark L (2003, October). *What is data mining and what are its issues*. Retrieved on September 10, 2004, from <http://www.darwinmag.com/read/100103/mining.html>

SAS, *Eddie Bauer Uses SAS to Tailor Customer Relationships*. Retrieved on September 8, 2004, from <http://www.sas.com/success/eddiebauer.html>

Two Crows Corporation, (1999). *Introduction to data mining and Knowledge Discovery, Third Edition*. Retrieved on August 12, 2004, from <http://www.twocrows.com/articles.htm>.

WEKA Classifiers web site. Retrieved on November 22, 2004, from
<http://www.oefai.at/~alexsee/WEKA/doc/weka.classifiers.rules.Ridor.html>

Whiting Rick, (2001, August 6). *Carrier Fans Sales with data mining*, Information Week.
Retrieved on September 9, 2004, from
<http://www.informationweek.com/showArticle.jhtml?articleID=6506443>.