Data Mining Technologies Inc.

Knowledge Discovery From Databases

A New Technology for Data Mining

White Paper

Michael Gilman, PhD

Data Mining Technologies Inc.
Melville, NY 11714
(631) 692-4400
e-mail mgilman@data-mine.com
www.data-mine.com

June 2006

Management Summary - The Bottom Line

Data Mining offers great potential in business and science because it can find valuable information that other methods can't.

This paper will provide an overview of data mining techniques and provide you with examples of how you can use it to achieve significant business goals or enhance your research effort.

This paper will describe how **Nuggets®,** the next step in data mining, gives you more power and ease of use to extract knowledge from your data and why this may be important to your organization.

Nuggets®, analyzes data by automatically formulating hypotheses about data and saving the correct ones as the "model". It performs all the necessary functions necessary to test and deploy the model in an easy to use program.

Data mining can achieve high return on investment decisions by exploiting one of an enterprise's most valuable and often overlooked assets—DATA!

Overview

Today's business environment is more competitive than ever. The difference between survival and defeat often rests on a thin edge of higher efficiency than the competition. This advantage is often the result of better information technology providing the basis for improved business decisions. The problem of how to make such business decisions is therefore crucial. One answer is through the better analysis of data. Data mining is a methodology to assess the value of the data and to leverage that value as an asset to provide large returns on the analytic investment.

The problem that often confronts researchers new to the field is that there are a variety of data mining techniques available—which one to choose? All these tools give you answers. Some are more difficult to use than others, and they differ in other, superficial ways, but most importantly, the underlying algorithms used differ and the nature of these algorithms is directly related to the quality of the results obtained and ease of use. This paper will address those issues.

Some estimates hold that the amount of information in the world doubles every twenty years. Undoubtedly the volume of computer data increases at a much faster rate. In 1989 the total number of databases in the world was estimated at five million, most of which were small local computer files. Today the automation of business transactions produces a deluge of data because even simple transactions like telephone calls, shopping trips, medical tests and consumer product warranty registrations are recorded in a computer. Scientific databases are also growing rapidly. NASA, for example, has more data than it can analyze. The 2000 US census data of over a billion bytes contains an untold quantity of hidden patterns that describe the lifestyles of the population. Sales transactions by people and businesses consume untold terabytes – and on it goes.

Most of this data will never be seen by human eyes and even if viewed could not be analyzed by "hand." But is this data valuable? The answer is usually an emphatic "yes".

Byte Magazine reported that some companies have reaped returns on investment of as much as 1,000 times their initial investment on a single project. More and more companies are realizing that the massive amounts of data that they have been collecting over the years can be their key to success. With the proliferation of data warehouses, this data can be mined to uncover the hidden nuggets of knowledge. Simply put, data mining tools are fast becoming a business necessity. The Gartner group has predicted that data mining will be one of the five hottest technologies in the early years of the new century.

There are currently several computer data mining techniques available. Not all of these are equal in effectiveness. This white paper will discuss the leading ones and an exciting and powerful new data mining method, Nuggets® that uses breakthrough technology to offer significant benefits over other methods.

What is Data Mining?

The objective of data mining is to extract valuable information from your data, to discover the "hidden gold." This gold is the valuable information in that data. Small changes in strategy, provided by the data mining discovery process, can translate into a difference of millions of dollars to the bottom line or better scientific decisions. With the proliferation of data warehouses, data mining tools are fast becoming a business necessity. An important point to remember, however, is that you do not need a data warehouse to successfully use data mining—all you need is data.

Many traditional reporting and query tools and statistical analysis systems use the term "data mining" in their product descriptions. Which leads to the question, "What is a data mining tool and what isn't?" The ultimate objective of data mining is knowledge discovery and data mining methodology is a technique to extracts predictive information from databases. With such a broad definition, however, an on-line analytical processing (OLAP) product or a statistical package could qualify as a data-mining tool, so some have narrowed the definition. A data mining method should unearth information *automatically*. By this definition data mining is data-driven, whereas by contrast, traditional statistical, OLAP, reporting and query tools are user-driven.

Nuggets® Methodology True Rule Induction

Nuggets® uses proprietary search algorithms to develop English "if - then" rules. The benefit of this is that results (i.e. predictions) are more understandable to the decision makers. Here's an example of the data mining rules that Nuggets® might discover for a project to target potential product buyers.

Rule 1.
IF CUSTOMER SINCE = 1998 through 2002
AND REVOLVING LIMIT = 5120 through 8900
AND CREDIT/DEBITRATIO =67
THEN Potential Buyer = Yes with a confidence factor of 89%

Rule 2.
IF CUSTOMER SINCE = 1998 through 2002
AND REVOLVING LIMIT = 1311 through 5120
AND CREDIT/DEBITRATIO = 67

THEN Potential Buyer = Yes with a confidence factor of 49%

Nuggets uses genetic methods and learning techniques to "intelligently" search for valid hypotheses that become rules. In the act of searching, the algorithms "learn" about the training data as they proceed. The result is a very fast and efficient search strategy that does not preclude any potential rule from being found. The new and proprietary aspects include the way in which hypotheses are created and the searching methods.

Nuggets® also provides a suite of tools to use the rules for prediction of new data, understanding, classifying and segmenting data. Additionally it provides functionality for data cleansing, sampling (including stratified sampling), binning and visualization.

When a model is created it can be tested through the use of the validation capabilities provided by Nuggets® on holdout or other data. Validation reports and graphs allow you to assess the "predictability of the data.

Nuggets® comes with an easy to use interface that makes creation of a model as easy as a few clicks without sacrificing power.

Pros
This method is fast and efficient in finding patterns. Tools are provided that allow you use of the rules to predict a file of new data, Nuggets® handles highly non-linear relationships and noisy or incomplete data. It can model large numbers of variables which is useful in modeling purchase transactions, click stream data or gene problems for example. Has artificial variable feature that not only predicts but tells you the optimum action to achieve that prediction. Numerous other unique features are provided.
Currently runs on Windows 98, Windows 2000, NT and XP although data can be imported from other platforms.

Cons
Does not run directly on mainframes but can import data to run on client PC's and work as a client-server system.

What Nuggets® is Not

Nuggets® is *not* a statistical tool. It does not require restrictive statistical assumptions such as independence, linear relationships, multi-colinearity, normality, etc. It finds rules for which a set of independent variables are correlated with a result. This non-statistical notion of correlation simply means that given the 'IF' condition, the 'THEN' condition occurs a given percentage of time.

For example suppose we develop the following rule:

IF Credit Rating = Good AND Bank Balance = over $10,000 And Employed = Yes Then Successful Loan = Yes, with confidence factor of 87%

This means that using the examples in the training file: of those, which satisfied the 'If' condition, 87% turned out to be successful. Thus the predictor variables, in this case credit rating, employment and bank balance, were correlated (i.e. associated) with a successful loan. Notice that Nuggets® is not suggesting a cause and effect relationship. A bank balance of over $10,000 is probably not the cause of the loan being good. It is merely associated with it in combination with the other factors as stated by the rule.

Nuggets® is a data mining system for PC users that puts the power of a complete data mining environment on everyone's desktop. It uses powerful new rule induction methodology to make explicit the relationships in both numeric and non-numeric information.

Nuggets® is automatic. This means that Nuggets® finds rules automatically.

Nuggets® then uses the rule library it has built to forecast expected results from new information, based on the "experience" contained in your existing database.

Other Data Analysis Methods - An Overview

The following represents a discussion of some of the most popular methods used to extract information from data.

Non-Data Mining Methods

Query Tools

Most of these tools come with graphical components. Some support a degree of multi-dimensionality such as crosstab reporting, time series analysis, drill down, slice and dice and pivoting.

Pros

These tools are sometimes a good adjunct to data mining tools in that they allow the analyst an opportunity to get a feel for the data. They can help to determine the quality of the data and which variables might be relevant for a data-mining project to follow. They are useful to further explore the results supplied by true data mining tools.

Cons

Simply put -- you must formulate the questions specifically. What are the sales in the northeast region by salesperson and product? If a person's income is between $50K and $100K what is the probability they will respond to our mailing? What percentage of patients will have nausea if they take penicillin and if they also take a beta-blocker?

This approach works well if you have the time to investigate the large number of questions that may be involved, which you almost never will. For example, a data-mining problem with 200 variables where each variable can have up to 200 values has 1.6×10^{460} values. This number is so large that all the computers on earth operating for the rest of the life of our galaxy could not search among these possibilities. Nuggets®, however, while not exploring them all *explicitly,* does examine them *implicitly* through **intelligent search methods** that avoid the insignificant ones.

Querying, therefore, is most effective when the investigation is limited to a relatively small number of "known" questions.

Data Mining Methods

Statistics

There are several statistical methods used in data mining projects that are widely used in science and industry and provide excellent features for describing and visualizing large chunks of data.

Some of the methods commonly used are regression analysis, correlation, Chaid analysis, hypothesis testing, and discriminant analysis.

Pros

Statistical analysis is sometimes a good 'first step' in understanding data. These methods deal well with numerical data where the underlying probability distributions of the data are known. They are not as good with nominal data such as "good", better", "best" or "Europe", "North America", "Asia" or "South America".

Cons

Statistical methods require statistical expertise, or a project person well versed in statistics who is heavily involved. Such methods require difficult to verify statistical assumptions and do not deal well with non-numerical data. They suffer from the "black box aversion syndrome". This means that that non-technical decision makers, those who will either accept or reject the results of the study, are often unwilling to make important decisions based on a technology that gives them answers but does not explain how it got the answers. To tell a non-statistician CEO that she or he must make a crucial business decision because of a favorable R statistic is not usually well received. With Nuggets® you can be told exactly how the conclusion was arrived at.

Another problem is that statistical methods are valid only if certain assumptions about the data are met. Some of these assumptions are: linear relationships between pairs of variables, non-multicollinearity, normal probability distributions, independence of samples. If you do not validate these assumptions because of time limitations or are not familiar with them, your analysis may be faulty and therefore your

results may not be valid. Even if you know about them you may not have the time or information to verify the assumptions.

Neural Nets

This is a popular technology, particularly in the financial community. This method was originally developed in the 1940's to model biological nervous systems in an attempt to mimic thought processes.

Pros

The end result of a Neural Net project is a mathematical model of the process. It deals primarily with numerical attributes but not as well with nominal data.

Cons

There is still much controversy regarding the efficacy of Neural Nets. One major objection to the method is that the development of a Neural Net model is partly an art and partly a science in that the results often depend on the individual who built the model. That is, the model form (called the network topology) and hence the results, may differ from one researcher to another for the same data. There is the problem that often occurs of "over fitting" that results in good prediction of the data used to build the model but bad results with new data.

The "black box syndrome" also applies here.

Decision Trees

Decision tree methods are techniques for partitioning a training file into a tree representation. The starting node is called the root node. Depending upon the results of a test this node is then partitioned into two or more sub-sets. Each node is then further partitioned until a tree is built. This tree can be mapped into a set of rules.

Pros

Fairly fast and results can be presented as rules.

Cons

By far the most important negative for decision trees is that they are forced to make decisions along the way based on limited information that implicitly leaves out of consideration the vast majority of potential rules in the training file. This approach may leave valuable rules undiscovered since decisions made early in the process will preclude some good rules from being discovered later.

How Nuggets® Can Help Your Organization

Features

- User friendly, intuitive interface
- Available for desktop or as on line real time data mining engine
- Power to extract knowledge from data that other methods can not
- Automatic rule generation in English "if-then" rules
- Ability to model up to 50,000 variables (without using clustering)
- Employs machine learning (No statistics used)
- Automatic binning of numeric variables
- Binning of nominal variables
- Ability to handle complex non-linear relationships with no statistical requirements
- Handles missing data
- Handles noisy data
- Assists in finding data errors
- Provides validation module
- Provides predictions for new data
- Reverse engineers information implicit in databases

- Allows stratified sampling for training files
- Unique feature that resolves rule conflicts for better predictions
- Computes attribute significance without limitation of correlated variables

Areas of Potential Application

The following list includes only a few of the possible applications.

Business

- CRM
- Banking -- mortgage approval, loan underwriting, fraud analysis and detection
- Finance -- analysis and forecasting of business performance, stock and bond analysis
- Insurance -- bankruptcy prediction, risk analysis, credit and collection models
- Web Marketing – personalization, targeted banner ads and cross sell/upsell opportunities
- Direct Marketing – response models, churn models, optimum creative, next to buy analysis
- Government – threat assessment, terrorist profiling
- Operations Improvement – find parameters that optimize efficiency, predict efficiency

Manufacturing

- Process, Improvement, fault analysis, quality control, preventive maintenance scheduling, automated systems

Medicine

- Gene analysis, epidemiological studies, toxicology, diagnosis, drug interactions, risk factor analysis, quality control, retrospective drug studies

Scientific Research

- General modeling of all types

Technical Information

Nuggets® is a true 32-bit system that will run on Windows 98, 2000, NT, XP. It can run on a standalone desktop or in such complex environments as client server and parallel processing enabled systems. It predicts, forecasts, generalizes and validates.