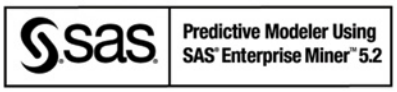
Business Intelligence
Solutions

Head Office
T: 267.616.1444
F: 1.905.761.1006
9 Flagstaff Place
Philadelphia, PA
19115 USA

info@bisolutions.us

Toronto
T: 647.271.1932
F: 905.761.1006
150 Borrows Street
Thornhill, ON
L4J 2W8 Canada

www.bisolutions.us



www.BISolutions.us | Data. Knowledge. Action.

WHITE PAPER:

DATA MINING: THE MEANS TO COMPETITIVE ADVANTAGE

APRIL 2008

Dmitry Brusilovsky & Eugene Brusilovskiy

Summary: By investing in data mining applications, an organization can gain a competitive advantage and uncover valuable customer information that cannot be identified in any other way.

Data mining technology is rapidly becoming strategically important to many data-rich firms. In fact, it is now considered a significant new component of the enterprise decision support system and this increase in demand is forcing managers to ask themselves the following questions (Brusilovskiy and Hernandez, 2001):

- Should we buy data mining software or not?
- Would it be best to hire data mining analysts, retrain current employees or invite an external data mining consultant to join our company?
- For which challenges is data mining technology most appropriate?

There are no universal answers to these questions. Rather, the answers depend on an organization's industry, business and project specificity. In order to respond with confidence, management must become familiar with the fundamental characteristics of data mining technology. They need to understand that a successful data mining project requires them to: a) select a problem that is appropriate for data mining, and b) select the most suitable data mining technology and/or data mining consultant for that problem.

WHAT CONSTITUTES A REAL DATA MINING PROJECT?

Data mining is defined as the non-trivial iterative process of extracting implicit, previously unknown and potentially useful information from data. The successful application of data mining requires the selection of a suitable project. The business problem that such a project may address can be labeled as structured or unstructured (Table 1).

Structured problems can be categorized as those for which all structural elements – such as goals, alternatives, data structure, criteria and environmental factors – are well known. At a glance, working on data mining projects that uncover previously unknown non-trivial hidden patterns in the data of structured problems may seem like a smart way to gain a competitive advantage. However, your competitors are most likely equally well acquainted with the known elements of such problems and can consequently match your company's results if they use corresponding analytics.

A greater competitive advantage lies in unstructured problems. Simplicity of pattern ascertainment, automation of hidden pattern discovery and minimal pattern interpretation issues do not apply to unstructured problems. By their nature, unstructured problems have no standard solutions (Table 1). Because some or all of the structural elements of an unstructured problem are undefined, ill-defined or unknown, these problems are difficult to solve.

Your competitors are not familiar with your company's unstructured problems. In fact, they're probably not even aware that such problems exist. Your real competitive edge lies in using knowledge-discovering technology to solve your business's unstructured problems, as your competitors will be unable to easily replicate your solutions.

In summary, projects that find the solutions to unstructured business problems can be termed real data mining projects. The successful resolution of a real data mining project provides organizations with high payoff.

TABLE 1: UNSTRUCTURED VS. STRUCTURED BUSINESS PROBLEMS

ATTRIBUTE	WELL-STRUCTURED BUSINESS PROBLEM	UNSTRUCTURED BUSINESS PROBLEM
Characteristics	Can be:	Cannot be:
	described with a high degree of completeness	described with a high degree of completeness
	solved with a high degree of certainty	resolved with a high degree of certainty
	easily and uniquely translated into quantitative counterpart	easily and uniquely translated into quantitative counterpart
	Experts usually agree on the best method and best solution	Experts usually disagree on the best method and best solution
Goal	Find the best solution	Find reasonable/acceptable solution
Complexity	From very simple to complex, usually lies within one discipline	From complex to very complex, as a rule lies at the interface of multiple disciplines
Causality	Easy to ascertain	Difficult to ascertain
Consequences of potential actions	Known	Unknown
Example	Project: Sample size identification	Project: Customer Feedback Study
Key business question	What is customers sample size (only one promotion channel vs. synergy of two promotional channels) to detect at least two purchase difference in sales of Product A?	Is there any relationship between customer perception of company reputation/brand equity and company products portfolio sales?
Translation into quantitative counterpart	If the project will be given to 5 different statisticians, all of them will come up with identical results: same methodology, same sample size calculation procedure, and same outcome – a table with calculated sample size as a function of power, regardless of software used	If the project will be given to 5 different data miners, each of them will come up with unique methodology, require specific data mining algorithms and software, and the results will not be identical (even terms in which findings will be formulated, can be completely different)

WHAT IS THE DIFFERENCE BETWEEN STATISTICAL ANALYSIS AND DATA MINING?

Statistical analysis is designed to deal with well structured problems. Results are software and researcher independent, and inferences reflect statistical hypothesis testing.

Data mining is designed to deal with unstructured problems. Results are software and researcher dependent, and inferences reflect the computational properties of the data mining algorithm at hand (Table 2).

TABLE 2: STATISTICS VS. DATA MINING: CONCEPTS

FEATURE	STATISTICS	DATA MINING
Type of problem	Well structured	Unstructured / Semi-structured
Inference role	Explicit inference	No explicit inference
Objective of the analysis and data collection	Objective formulation, followed by data collection	Data rarely collected for the objective of analysis/modeling
Data set	Data set is small and hopefully homogeneous	Data set is large and heterogeneous
Paradigm/approach	Theory-based (deductive)	Synergy of theory-based and heuristic-based approaches (inductive)
Signal-to-noise ratio	STNR > 3	0 < STNR <= 3
Type of analysis	Confirmative	Explorative

Both statistical and data mining paradigms provide a wide range of regression methods/models. Modern statistical regression models include but are not limited to: robust regression, quantile regression, loess/kernel regression, additive regression and spline regression. Data mining regression models include tree-based regression (CART, CHAID, and others), TreeNet, Random Forrest, Multilayer Perceptron, MARS and others.

What is the difference between statistical and data mining regression? In short, the difference lies in the varying degrees of tolerance that statistical and data mining models have to viral data anomalies (Table 3). Data mining approaches can handle high-dimensional heterogeneous data with a high degree of sparseness and multicollinearity, and with a significant percentage of outliers/leverage points and missing values, and are able to discover uncharacterizable non-linearities among differently scaled variables in high-dimensional space. Statistical approaches cannot (Brusilovskiy 2007).

TABLE 3: STATISTICS VS. DATA MINING: REGRESSION MODELING

FEATURE	STATISTICS	DATA MINING
Number of inputs	Small	Large
Number of observations	Small	Large
Type of inputs	Interval scaled and categorical with small number of categories (percentage of categorical variables is small)	Any mixture of interval scaled, categorical and text variables
Multicollinearity	Wide range of degree of multicollinearity with intolerance to high degree of multicollinearity	Severe multicollinearity is always there, tolerance to any degree of multicollinearity
Distributional assumptions, homoscedasticity, outliers, missing values	Intolerance	Tolerance
Type of model	Linear / non-linear / parametric / non-parametric in low dimensional X-space (intolerance to uncharacterizable non-linearities and sparseness)	Non-linear and non-parametric in high dimensional X-space with tolerance to sparseness and uncharacterizable non-linearities

WHEN IS DATA MINING TECHNOLOGY APPROPRIATE?

Data mining technology is appropriate when:

- the business problem is unstructured
- the data includes a mixture of interval, nominal, ordinal, count, and text variables, and the role and number of non-numeric variables are essential
- the set of inputs includes many irrelevant and redundant variables
- the relationship among variables could be non-linear with uncharacterizable nonlinearities
- the data is highly heterogeneous with a large percentage of outliers, leverage points and missing values
- the sample size and the number of variables are both large

WHAT ARE THE KEY CHARACTERISTICS OF DATA MINING APPLICATIONS?

UNIQUENESS

Even if a number of companies appear to face the same business challenges each will require its own unique data mining application. Workers compensation health care projects, for example, often use available claims data to understand and answer questions about cost effectiveness and quality medical/litigation/disability management outcomes. However, the difference between the data structures and performance measurements of each unique company, as well as the difference in questions each company formulates, will determine the appropriate analysis and methods for gathering reliable information for effective decision-making.

Say three organizations share a similar business challenge. While one might assume that the data mining project required for each challenge would be similar, this is not necessarily the case. One may only need claims data to predict the costs and duration of a claim for a case assessment goal. Another may need to use benchmark information to identify the best clinical/litigation practices. Yet another may require both safety and claims data to determine the effectiveness of safety prevention programs.

The methodology of analysis can change dramatically from company to company, and from project to project, depending on the strategies, goals, data, available software, analytical tools and organizational culture.

ABSENCE OF IMPLEMENTATION STANDARDS

Because unstructured problems are complex, results may be unpredictable. A data miner will often not know how hidden patterns can be identified and described until the data mining analysis problem has been formulated and different data mining tools have been applied.

Take the following example: a data mining analyst was asked to determine whether safety prevention programs have a significant effect on personal injury risk reduction. The data miner tested this hypothesis using verification-driven data mining and found no relationship between the two. Formulating and exploring other hypotheses, however, the data miner found that the roles of safety programs and personal injury risk reduction were reversed. Personal injuries were the driver for more safety programs and disciplinary action. Thus, the company's safety prevention programs were reactive instead of proactive.

NEED FOR A MULTIDISCIPLINARY TEAM

A basic team consists of an executive sponsor, a quantitative professional (M.S. or Ph.D. in applied statistics, operations research, artificial intelligence or decision science), a data management specialist, a subject matter person(s), and a project manager with a wide range of business, IT and quantitative experience. On occasion, a successful data mining project is implemented by one person who takes on the role of business expert, data expert and analytical expert simultaneously.

Many unstructured problems are cross-functional by nature. A data miner may uncover significant results but be unable to act on those results if he or she does not have the participation and support of those key decision-makers with the authority to approve and implement actions based on the results. Decisions often require a change in how an area is managed and without the participation of management, no commitment to change exists.

SYNERGY OF DATA MINING METHODS

Data mining can be divided into two classes – verification driven and discovery driven. The first class is associated with traditional quantitative approaches. The second class is induced with knowledge discovery technology and involves discovery of previously unknown patterns hidden in the data.

Verification-driven data mining allows decision-makers to express and verify organizational and personal domain knowledge. This methodology consists of developing hypotheses and determining whether they are acceptable by applying various statistical and other quantitative methods to available data and subject matter judgment. Verification-driven and discovery-driven data mining creates a synergy that produces more meaningful, more interpretable and more reliable results with less iteration.

SOFTWARE DEPENDENCY

The data mining umbrella covers different architecture neural networks (Bayesian belief network, multilayer perceptron, radial basis function network, etc.), decision tree algorithms, logistic regression, fuzzy logic, genetic algorithms, clustering, associations analysis, memory based reasoning, and other analytical methods. No software company offers all of the above-mentioned data mining methods.

For any unstructured problem experts often disagree about the best method and best solution, and the problem cannot be uniquely translated into its quantitative counterpart (Table 1). Since data mining is a set of heuristic-based methods (Table 2), each data mining software vendor tries to introduce its own heuristics to improve the data mining algorithm. Therefore, even if two different software vendors offer the same data mining methods, they may have slightly different algorithms due to the company's heuristics or the use of different optimization algorithms and/or different random number generators. Along with this, for the same data mining method each vendor provides a different set of options that can be selected by data miner. Taking into account the inherent instability of some data mining algorithms, it is very unlikely that identical results will be produced when the same data mining method but several different vendors' data mining software applications are employed.

THE ESSENTIAL ROLE OF CREATIVITY

Solving unstructured problems almost appears to be more art than science. A sound solution requires the data miner to possess systemic insight of the problem under consideration, good intuition, the ability to generate dissimilar heuristics and experience in dealing with key decision makers who do not completely understand the data-mined information, especially if it contradicts their dominant assumptions.

Creative solutions typically have an inherent risk due to the complexity of the problem and the need to journey through unknown territory. We all find comfort when we apply familiar solutions to problems and stick to what we know best. Sometimes the solutions are quite visible, but very often they are not. If the solution were obvious to everyone or easy to see, it would likely have been found already. Finding solutions in the darkness and creating nontrivial knowledge are dynamic and courageous processes.

HOW TO SELECT A DATA MINING CONSULTANT?

Source: King, E. A. (2005), How to Buy Data Mining: A Framework for Avoiding Costly Project Pitfalls in Predictive Analytics, DM Review Magazine, October 2005 issue

“One misconception about selecting an external data mining consultant is that the consultant should also be an expert in your industry. It may be helpful for the consultant to have background in order to speak and interpret industry lingo while appreciating the competitive environment and primary drivers, but unlike building a knowledge base, it is actually preferable not to have the industry's strongest domain expert who also happens to do some data mining. While the consultant may appear impressive at the outset, too much industry expertise can introduce subjectivity and preconceived notions that may skew the way models are developed and interpreted.

Models by their nature are objective, and the consultant should be, too. The best results are achieved when the data mining expert drives the model-building process but not the results. The data mining consultant should then work with your organization's domain expert to mutually interpret the results, validate them and determine the most effective way to make them useful.”

CONCLUSION

Digging for a sustainable competitive advantage requires companies to focus on real data mining projects and to build a team with the right synergy to discover meaningful results. This is not an overnight process. It would be a wise management decision to start using data mining technology as soon as possible, because gaining sustainable company knowledge is a recursive process and data mining technology is a necessary component to remain competitive. Understanding the nature and characteristics of a real data mining study will better prepare management to select and handle the right data mining projects that will offer their company a competitive edge.

REFERENCES

1. Kumar, T. *An Introduction to Data Mining in Institutional Research* http://www.airweb.org/webrecordings/Kumar%20Data%20Mining%20Workshop.pdf
2. Hand, D. J. *Data Mining: Statistics and More? The American Statistician, May 1998, Vol. 52 No. 2* http://www.amstat.org/publications/tas/hand.pdf
3. Friedman, J.H. 1997. *Data Mining and Statistics. What's connection? Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas*
4. Smyth, P. (2000), *An Introduction to Data Mining, Elumetric.com Inc*
5. King, E. A. (2005), *How to Buy Data Mining: A Framework for Avoiding Costly Project Pitfalls in Predictive Analytics, DM Review Magazine, October 2005 issue* http://www.dmreview.com/article_sub.cfm?articleId=1038094
6. Brusilovskiy, P. (2007), *Data Mining in Pharmaceutical Marketing and Sales Analysis. 2007 ICSA Applied Statistics Symposium* statgen.ncsu.edu/icsa2007/talks/Session%203D%20Data%20Mining%20in%20Pharmaceutical%20Marketing%20.ppt
7. Brusilovskiy, P. and by Hernandez, R. (2001), *Data Mining for a Sustainable Competitive Advantage. DM Review Magazine, June 2001 issue* http://www.dmreview.com/editorial/dmreview/print_action.cfm?articleId=3508

