

Pharmaceutical Application of SAS® Enterprise Miner™

Thomas H. Burger, Eli Lilly and Company, Indianapolis, IN

Jerry E. Plaatje, SAS Institute Inc., Cincinnati, OH

John M. LaBore, Eli Lilly and Company, Indianapolis, IN

James K. Malone, Eli Lilly and Company, Indianapolis, IN

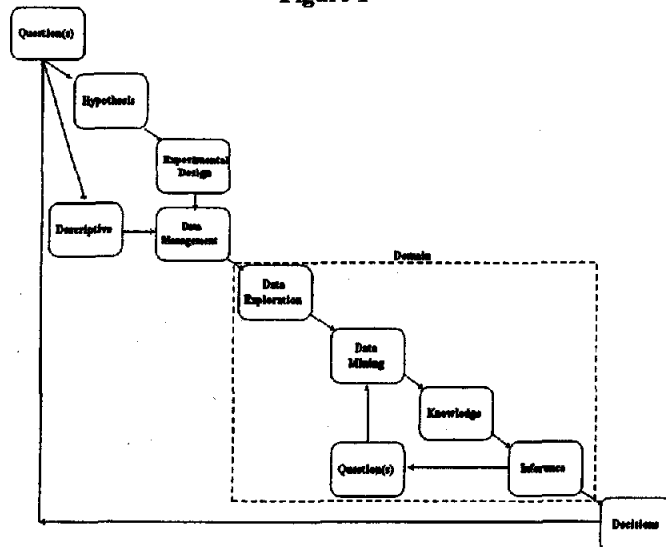
Abstract

Competitive advantage requires rapid formation of sound inferences for product decisions. As information repositories, warehouses store teradata for decision support. Strategic use of this data requires techniques proven for information delivery. Mining reveals relationships for knowledge formation through methodological data selection, exploration and modeling. A structured process ensures conformance to business assumptions for reliable interpretation. To illustrate, we discuss data mining in pharmaceutical research.

Introduction

The pharmaceutical industry is information driven and highly competitive. Regulation mandates that decisions rest on a results-bound basis. Rigorous information delivery is required for sound inferences of safe and effective disease treatment (Figure 1). This demands integrated tools that effectively couple technology with business needs in the domain of data analysis and exploration.

Figure 1



Data warehouses assimilate information for analysis and presentation. *Interpretation* of data is key to their effective use. While standard analytical methods are appropriate for submissions, recognition of unknown relationships requires techniques for data exploration, mining and meta-analysis. *Exploration* identifies relationships warranting (deeper) analysis; *mining* enables knowledge formation through modeling and algorithmic searches for pattern. It therefore provides a useful mechanism for exploration and generation of new questions. Commonly used for analysis of sales and marketing data at the enterprise level, mining is also ideal for other arenas. Analytical questions abound in early development and manufacturing R&D. Data in these areas are typically balanced, complete and have great potential for

knowledge discovery through mining for interpretation, inference formation and business application.

Data mining has emerged as key to information discovery by providing a mechanism for identifying unknown relationships. Berry and Linoff (1997) describe a *Virtuous Cycle of Mining*; SAS Institute advocates a systematic framework using a flexible iterative stepwise SEMMA process coupled to Enterprise Miner.

Problem Statement

Pharmaceutical product development is a complex process requiring methods for efficient knowledge formation. Identification of relationships through exploration and interpretation produces sound inferences that speed product delivery. However, the volume and complexity of information demands sophisticated techniques for data assimilation and interpretation. Recent advances in technology and methodology provide mining tools for this purpose.

Mining is widely used in sales and marketing research to interpret incomplete indicators and identify unknown relationships. Previously restricted to those with programming expertise, SAS Enterprise Miner™ (EM) now enables non-programmers to mine. While sophistication permits point-and-click multi-tier analyses, *its use should be guided by statistical and business experts.*

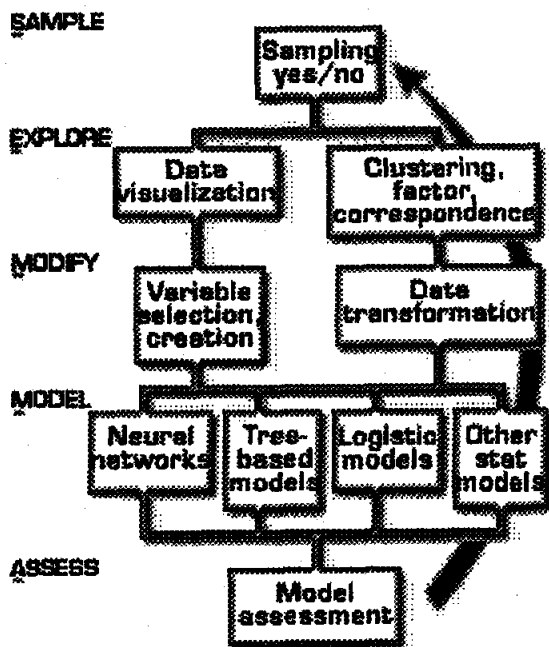
Objective

Data generated in early development and manufacturing typically have a very *complete* structure due to control of design, data generation and acquisition. Interpretation of these data through analysis enables sound hypothesis and inference formation. Mining enables enterprise-level knowledge discovery through exploration and meta-analysis to aid submission decisions.

Mining enables exploration, analysis and presentation for discovering relationships among the multidimensional longitudinal variables common to these parts of the business. These discoveries increase organizational intelligence by tying analytical techniques to the business for better decision making.

Methodology

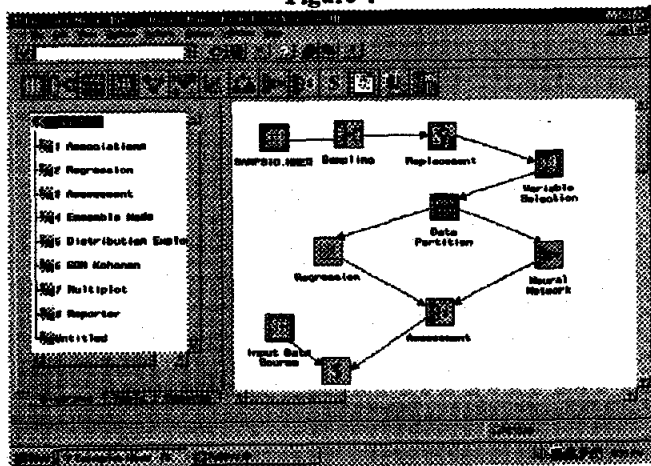
Mining is the exploration and analysis of data to discover unknown relationships. It can be performed in a top-down fashion for *hypothesis testing* or using a bottom-up approach for *knowledge discovery*. In both cases, it is essential to tightly couple mining with its business context. The top-down approach enables tests of hypotheses about relationships. Special consideration is required to ensure hypotheses conform to the allowable inferential space of data. The bottom-up



Copyright SAS Institute, Inc

Enterprise Miner™ (EM) is a stand-alone integrated SAS tool enabling a broad range of exploration, modeling and graphical techniques. In addition to sophisticated analytical modeling, it provides tools for identifying logical approaches to data exploration (Figure 4). Intended for use with the SEMMA process, it provides end-to-end mining solutions too numerous to describe here.

Figure 4



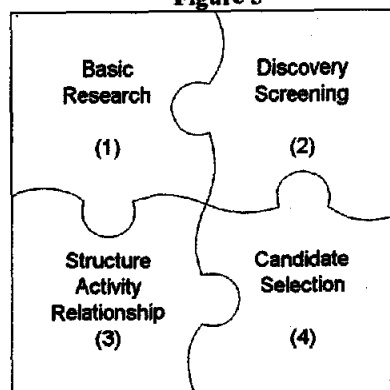
Copyright SAS Institute, Inc.

Application

Mining has traditionally been applied programmatically; however, recent technology has simplified the task. Widely used in sales and marketing (Berry and Linoff 1997, Rud 1999), mining's current pharmaceutical use is limited. Cohen and Rud (1999) recommend mining for sales, marketing and other industry sectors. We propose non-traditional use for it in non-GxP R&D.

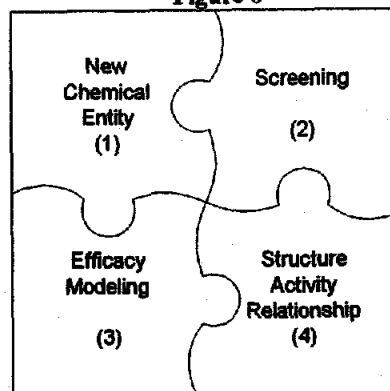
Whereas GxP analyses require standard approaches to system Design (Burger and Pochon 1997, Burger et al. 1999), Validation (Pochon and Burger, 1999) and Warehousing (Pochon and Burger, 2000), early development is less constrained. Early discovery is comprised of data-rich components with balanced, complete data sources (Figure 5). Common to these areas is the need to prepare (Burger et al., 2000) and interpret huge volumes of data for discovery research. Exploration reveals unrecognized relationships and improves hypothesis testing to aid inference formation (Burger and Burger, 2000). This speeds knowledge discovery by identifying strong leads earlier and dropping false leads sooner fostering intelligent business decisions.

Figure 5



Discovery manufacturing discloses techniques for cost-effective compound formulation (Figure 6). Frequently, data sources are not only balanced and complete but also possess a range of process metrics. Mining this data may deepen understanding of the relationship of efficacy, stability and formulation for increased cost-efficiency.

Figure 6



Mining has broad applicability to other aspects of the business, too. It facilitates meta-analysis of cross-study trends, particularly since SAS Warehouse Administrator integrates with EM. Diseases could be better understood through retrospective analyses of historical clinical data for non-submission purposes. This could improve design and analysis of future compounds and clinical trials. Interpretation of Health Outcomes data through mining could also enhance understanding of therapeutic benefit through cross-patient profiles.

Summary

approach uses directed methods to predict and explain known relationships whereas undirected methods identify unknown relationships and recognize patterns. In directed mining, the model's output is specified prior to its creation, whereas undirected methods use the model to determine output patterns (Berry and Linoff, 1997).

Berry and Linoff (1997) describe the following steps as the *Virtuous Cycle of Mining*:

- 1) Identify Problem
- 2) Perform Mining
- 3) React
- 4) Measure Outcome

Note that (1) and (3) require business expert guidance.

Problem Identification begins with recognition of the business opportunity to be addressed through mining. It involves specification of the information required to reach conclusions and a plan for how this is to be accomplished.

Mining is the activity of exploring data for deeper understanding. Once begun, new information frequently leads to additional unplanned activities. The complexity of this process requires flexible tools specifically designed for exploration, modeling and visualization. SAS Enterprise Miner™ was expressly developed for this purpose.

Reaction is the process of incorporating knowledge discovery into business practices for improved decision making. Without this critical step, mining has limited utility.

Measuring Outcome is essential to assess the benefit mining provides in guiding process improvement to the business. The objective is to seek the desired response by asking the right questions.

Techniques

Mining typically consists of activities for Classification, Estimation, Prediction, Affinity Grouping, Clustering and Description. This brief discussion is adapted from Berry and Linoff (1997).

Classification, the most common technique, is used to construct models that group or categorize discrete elements based on some criteria. Decision Trees, Memory-Based Reasoning and Link Analysis are techniques suited to this method.

Estimation is used to generate values for unknown continuous variables, often for classification. Neural Networks are well suited to this method.

Prediction forecasts future values; used with historical data, it is similar to Classification and Estimation techniques. Market Basket Analysis, Memory-Based Reasoning, Decision Trees and Neural Networks are suited to this method.

Affinity Grouping is used for establishing associativity to generate rules from data. These techniques are broadly useful where transaction data is available, but knowledge of the pattern to search for is unknown.

Clustering is used to separate members of a heterogeneous population into homogenous clusters based on similarity; it may be used to preclude other mining techniques. It differs from classification in that it does not rely on predefined classes.

Description is used to characterize items to better enable an understanding of basic relationships. It is often used as a basis in explaining deeper relatedness.

Market Basket Analysis (MBA) uses clustering models to estimate the likelihood that items group. *Memory-Based Reasoning* (MBR) is a directed method using nearest-neighbor analysis to predict unknowns. *Cluster Detection* (CD) is an undirected method for detecting similarity. *Link Analysis* (LA) develops models based on pattern. *Decision Trees* (DT) are a directed methodology for classification. *Neural Networks* (NN) graphically model patterns for prediction, classification and clustering. *Genetic Algorithms* (GA) use directed methods on parameters to describe predictive functions. The relationship of these methods to the techniques described above is shown in Figure 2 (Berry and Linoff, 1997).

Figure 2

Technique	Class	Est	Pred	Aff Grp	Clus	Desc
Statistics	✓	✓	✓	✓	✓	✓
MBA			✓	✓	✓	✓
MBR	✓		✓	✓	✓	
CD					✓	
LA	✓		✓	✓		
DT	✓		✓		✓	✓
NN	✓		✓		✓	
GA	✓		✓			

Berry and Linoff (1997)

Process

The SEMMA framework consists of 5 steps for identifying logical relationships in data (Figure 3). The *Sample* step permits data subsetting to increase mining efficiency. The *Explore* step provides for data exploration and visualization. The *Modify* step allows alteration of data associations to improve modeling. The *Model* step enables a variety of modeling techniques. The *Assess* step permits cross-model comparisons of predictions.

Figure 3

Little has been published on the application of mining to pharmaceutical development. Many R&D business units generate data well suited to mining. Mining supplies powerful techniques for discovering unknown relationships and hypothesis testing for sound inference formation. SAS Enterprise Miner™ is effective using the methodologies we describe and can be a powerful tool for information delivery.

References

Berry, Michael J.A. and Gordon Linoff. 1997. *Data Mining Techniques: for Marketing, Sales, and Customer Support*. Wiley Computer Publishing. New York.

Burger, Thomas H. and Loren W. Burger. 2000. *Exploratory Analysis for Efficient Data Mining*. These proceedings.

Burger, Thomas H. and Philip M. Pochon. 1997. *Structured Design and CASE Tool Use in SAS® Macro Systems*. Proc. of the Eighth Annual MidWest SAS Users Group Conference. Chicago, IL.

Burger, Thomas H., Richard W. Tucker and John M. LaBore. 1999. *Design and Deployment of a Web-Application Using SAS/IntrNet*. Proc. of the Twenty-Fourth Annual SAS Users Group International Conference. Miami, FL.

Burger, Thomas H., Jerry Plaatje and John M. LaBore.
2000. *Removing Slag to Get the Ore: Preparing Data for Mining*. Proc of the 1st Annual Data Mining SAS Users group.
San Francisco, CA. (*In prep*)

Cohen, John J. and C. Olivia Rud. 1999. *Data Mining in The Pharmaceutical Industry*. Proc. of the 1999 Pharmaceutical SAS Users Group. New Orleans, LA.

LaBore, John M. and Thomas H. Burger. 1999. *Positioning SAS® for Corporate Effectiveness*. Proc. of the 1999 Pharmaceutical SAS Users Group. New Orleans, LA.

LaBore, John M. and Thomas H. Burger. 1999. *Integrating SAS® in Corporate IT*. Proc. of the 2000 Pharmaceutical SAS Users Group. Seattle, WA.

Pochon, Philip M. and Thomas H. Burger. 1998. *Validation of SAS® Macro Systems*. Proc. of the 1998 Pharmaceutical SAS Users Group. San Francisco, CA.

Pochon, Philip M. and Thomas H. Burger. 2000. *Warehousing, Metadata, and Object-Based Analysis*. Proc. of the Twenty-Fifth Annual SAS Users Group International Conference. Indianapolis, IN.

Rudd, C. Olivia. 1999. *Data Mining: An Overview of methods and Techniques for Increasing Profits in Direct Marketing*. Proc. of the Twenty-Fourth Annual SAS Users Group International Conference. Miami, FL.

Welbrock, P.R. 1998 *Strategic Data Warehousing Principles Using SAS Software*. SAS Institute Inc, Cary, NC. 384 pp.

Trademark Notice

SAS is a registered trademark of the SAS Institute Inc, Cary, NC and other countries.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Author Contact

Thomas H. Burger
Eli Lilly and Company
Lilly Research Laboratories
P.O. Box 708, GL43
150
Greenfield, IN 46140
Phone 317.277.7266

John M. LaBore
Eli Lilly and Company
Lilly Corporate Center
Drop Code 6334
Indianapolis, IN 46285
Phone 317.277.6387

Jerry E. Plaatje
SAS Institute Inc.
Cincinnati Regional Office
4555 Lake Forest Dr, Suite

Cincinnati, OH 45242
Phone 513.563.9444

James K. Malone
Eli Lilly and Company
Lilly Corporate Center
Drop Code 2042
Indianapolis, IN 46285
Phone 317.277.6472