

Removing Slag to Get the Ore: Preparing Data for Mining

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

Abstract

Competitive advantage requires rapid formation of sound inferences for knowledge discovery. Key to information delivery is data analysis and presentation. Warehousing enables storage of teradata for decision support; strategic use of this data requires proven techniques. Mining fosters disclosure of relationships through exploration and modeling. However, preliminary data preparation is essential to ensure it is performed upon data in its final desired form. Typically, steps are first required for data assimilation to identify and remove errors. Standard methods for data management increase reliability and speed this process. We discuss issues associated with preparation of data for mining.

Introduction

The data processing lifecycle is composed of multiple phases, each requiring specialized tasks (Figure 1). Following acquisition, warehouses enable data management, exploratory analysis and mining. Interpretation of this information relies upon both sound methodology and data integrity.

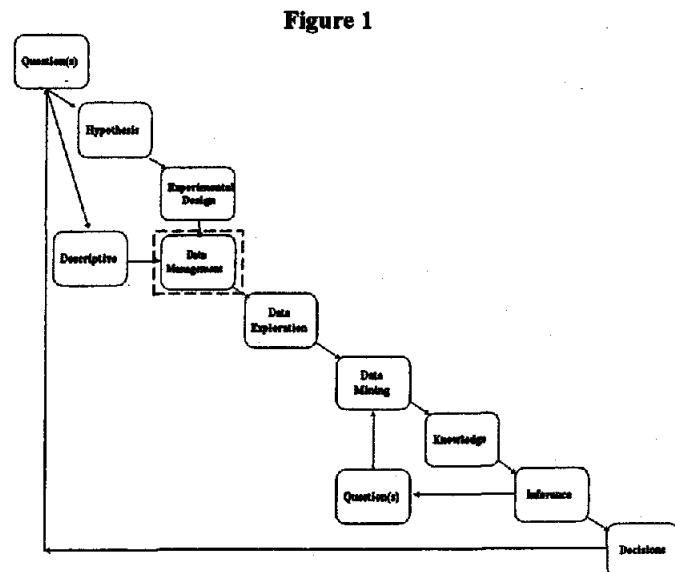


Figure 1

SAS offers multiple techniques for data management. Generally, this includes activities such as data access, merges, concatenation and integrity checks to identify discrepancies, errors and missing values. *Exploration* identifies relationships warranting (deeper) analysis; *mining* enables discovery of relationships through hypothesis testing, modeling and pattern searches. Such approaches provide a useful mechanism for decision support and generating business intelligence.

We discuss techniques applied to source data that are helpful to cleanse (errors), identify disparities (heterogeneity) and explore irregularities (outliers). We also briefly discuss use of the SAS Warehouse Administrator for data management.

Warehousing

Warehouses assimilate data for analysis and presentation. The precision, accuracy and reliability of data acquisition systems influence the extent of data preparation required. Physical data structure can also significantly impact the degree of manipulation necessary to establish desired relationships.

Data Preparation

Data management requires tools for manipulation and error removal. A fictitious data set with nine variables is used to demonstrate examples of data cleansing techniques (Figure 2).

Figure 2

Variable	Type	Length	Format	Label
CUSTOMER	Char	8		CUSTOMER NUMBER
NAME	Char	72	\$20.	CUSTOMER NAME
ADDR1	Char	32	\$20.	Address
CITY	Char	20	\$20.	City
STATE	Char	2	\$2.	State
ZIP	Char	5	\$5.	Zip code
REGION	Char	10		Region
INV	Char	8		INVOICE NUMBER
QTY	Char	8		Quantity purchased

A first pass of the data using the frequency procedure to examine the variable Quantity (QTY) shows the occurrence of several values having low frequency counts relative to other values of the variable, possibly indicating data errors (Figure 3). The output also provides a count of missing values. This procedure could be used to confirm uniqueness for a variable like Invoice Number.

Figure 3

The FREQ Procedure

Frequency Distribution				
Category	Frequency	Percent	Expected	Expected %
1	1780	49.42	1780	49.42
2	223	6.07	1821	78.60
3	277	7.57	2264	89.60
4	287	7.88	2821	90.64
5	1	0.03	1522	92.60
6	1	0.03	2523	94.92
7	1	0.03	1524	92.94
Total	1	0.04	1525	101.00

Frequency Missing = 4

Proc MEANS or UNIVARIATE may be used to examine data for potential errors. Proc MEANS will identify minimum and maximum values to allow you to make corrections based on knowledge of appropriate ranges. Proc UNIVARIATE reports several summary tables useful for identifying potential data errors. For example, the Extreme Observations table quickly points to values outside of normal (Figure 4). A graphic for this task may also be helpful and can be provided by SAS/Insight.

Figure 4

Extreme Observations			
Category	Value	Frequency	Value
1	2526	4	2519
1	2525	23	1871
1	2524	45	902
1	2521	233	846
1	2520	999	1744

Among SAS tools for data management and error-checking are:

- a DATA step with BY statement using a subsetting IF
- DATA step statements (such as MERGE, MODIFY, SET, SELECT, and UPDATE) with selected options
- the FORMAT procedure
- the FSEDIT procedure
- the FSBROWSE procedure with a WHERE statement
- the FSVIEW procedure with a WHERE statement
- the GCHART procedure
- the GPLOT procedure
- the REPORT procedure
- SAS/INSIGHT software
- the SHEWHART procedure
- the SQL Query window
- the SUMMARY procedure
- the TABULATE procedure

- the SORT procedure with the NODUPPLICATES or NODUPKEY option.

When performing data modifications, consider how SAS functions can help. This section provides a description of several useful functions with implementation examples. Using SAS functions you can calculate sample statistics, create SAS date values, convert zip codes to state postal codes, round values, generate random numbers, extract a portion of a character value, and convert data from one data type to another.

For example, the INDEX, INDEXC and INDEXW functions may be used to search for occurrence of character strings. This code will locate those observations that contain P.O. box numbers rather than street addresses. Note that the INDEX functions are case sensitive; so, alternatives are to use INDEX (upcase(addr1),'PO') or INDEX (lowcase(addr1),'po').

```
data postbox;
set example.rawdata;
if index (addr1,'PO') or
index (addr1,'P.O.') or
index (addr1,'POB') >0;
```

The SUBSTR function can be used to extract a portion of a value by starting at a specified location; while the SCAN function can be used to extract a portion of a value that is marked by delimiters. The Name variable of our example data set contains names in the form (Lastname, Title, Firstname). Using SCAN we can extract the first and last names into separate variables by using the SCAN function arguments of (variable name, word position, delimiters). The SUBSTR function might be used to extract only the last name as its position is always in the first character field.

```
data names;
set example.rawdata;
fname=scan(name,3,'. ');
lname=scan(name,1,'. ');
```

Similarly, variables may be combined using the TRIM function with concatenation. For example, the following code will create

```
data combo;
set names;
fname=trim(fname) || ' ' || trim(lname);
```

the fname variable containing the first and last names, stripped of trailing blanks, (which may be part of the Fname or Lname variable definitions) concatenated with a single blank delimiter.

The SPEDIS function is used to determine the likelihood of two character variable values matching expressed as the asymmetric spelling distance between the two values. The following code will complete every possible comparison of values using the variable selected (name) and create a data set that contains

those observations that meet a particular bracketed range. The values can then be examined for the likelihood of errors and corrected accordingly.

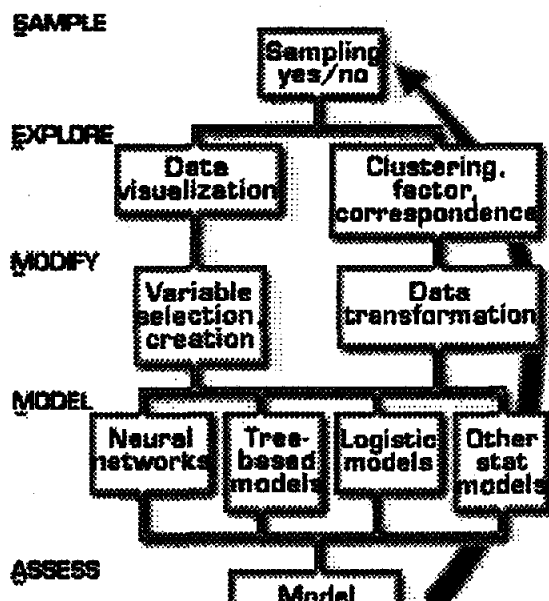
```
data dataset;
  set example.rawdata (keep=name) nobs=count;
  nname=name;
  do i=1 to count;
    set example.rawdata (keep=name) point=i;
    dist=spedis(nname,name);
    if 0 < dist < 10 then output;
  end;
run;
```

Data can be standardized programmatically. Sophisticated programs can be written that analyze data and make corrections without user intervention. The methodology of such programs can use techniques described above. In the variable *name* example, one might scan the name for the existence of a title, a multi-part last or first name, a lineage number or professional degree.

The SCAN function can be used to read and separate delimited strings of a character value into separate tokens. Individual tokens may then be analyzed to determined content. Where possible, tokens derived from a character string may be tested against a table of conversion values. Conversion tables include all possible values and the corresponding change value. For example, while standardizing an address variable, all variations on the spelling of the word 'street' would be converted to one consistent, standard form. The conversion table must, therefore, be very complete in listing every possible version of the word 'street' as well as 'avenue', road, lane, etc. Exception reporting is commonly used to identify values that have no match to the existing conversion table. Additionally, TRANSLATE and VERIFY can be used for character string manipulation.

Data Mining

Mining is the activity of exploring data for deeper understanding. 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 to business context. Special consideration is required to ensure hypotheses conform to the allowable inferential space of data.



SAS Institute advocates a systematic framework using a flexible iterative stepwise SEMMA process coupled to Enterprise Miner.

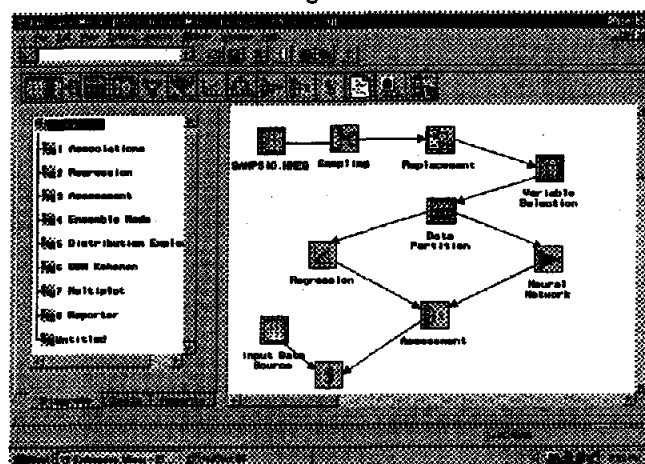
The SEMMA framework consists of 5 steps for identifying logical relationships in data (Figure 5). The *Sample* step permits data subsetting to increase mining efficiency. The *Explore* step provides for data exploration and visualization. The *Modify* step allows alteration of data associations to improve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

Figure 5
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 6).

Mining was previously restricted to those with programming expertise; 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*. Intended for use with the SEMMA process, the Enterprise Miner provides end-to-end mining solutions too numerous to describe here. An attractive feature of EM is that it is compatible with SAS Warehouse Administrator™ and JMP®. This greatly facilitates activities for table and metadata management during data preparation, exploration, visualization and mining.

Figure 6



Copyright SAS Institute, Inc

Summary

Much has been published on mining's application to warehoused data for knowledge discovery. It supplies powerful techniques for identifying relationships and hypothesis testing. SAS Enterprise Miner™ provides powerful tools for such activities. However, sound interpretation of data relies upon both sound methodology and data integrity. We describe standard techniques to prepare data for mining reliable 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*. Proc. of the 2000 Pharmaceutical SAS Users Group. Seattle, WA.

Burger, Thomas H., Jerry Plaatje, John M. LaBore and James K. Malone. 2000. *Pharmaceutical Application of SAS® Enterprise Miner*. Proc. of the 2000 Pharmaceutical SAS Users Group. Seattle, WA

Burger, Thomas H. and Philip M. Pochon. 1997. *Techniques for Warehousing Object-Based Analyses*. These Proceedings.

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.

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

Jerry E. Plaatje
SAS Institute Inc.
Cincinnati Regional Office
4555 Lake Forest Dr, Ste
Cincinnati, OH 45242
Phone 513.563.9444

John M. LaBore
Eli Lilly and Company
Lilly Corporate Center
Drop Code 6334
Indianapolis, IN 46285
Phone 317 277.6387