ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Data Mining Using SAS Enterprise Miner – A Case Study Approach

Kannan Subramanian[*1], Suresh Babu .G.N.K[2]

Assistant Professor, Dept. of MCA, Bharath University, Chennai, Tamil Nadu, India [1]

Head, Dept of MCA, GKM College of Engg, Vandular, Chennai, India[2]

[*]Corresponding Author

ABSTRACT: This document defines *data mining* as advanced methods for exploring and modeling relationships in large amounts of data. Your data often comes from several different sources, and combining information from these different sources may present quite a challenge. The need for better and quicker access to information has generated a great deal of interest in building data warehouses that are able to quickly assemble and deliver the needed information in usable form. To download documentation that discusses the Enterprise Miner add-ins to SAS/Warehouse Administrator, go to the SAS Customer Support Center *Web site* (***http://support.sas.com***). From Software Downloads, select ***Product and Solution Updates***. From the Demos and Downloads page, select ***SAS/Warehouse Administrator Software***, and download the version that you want. A typical data set has many thousand observations. An observation may represent an entity such as an individual customer, a specific transaction, or a certain household. Variables in the data set contain specific information such as demographic information, sales history, or financial information for each observation. How this information is used depends on the research question of interest. Ordinal variables may be treated as nominal variables, if you are not interested in the ordering of the levels. However, nominal variables cannot be treated as ordinal variables since there is no implied ordering by definition. To obtain a meaningful analysis, you must construct an appropriate data set and specify the correct measurement level for each of the variables.

KEYWORDS: **Interval** — a variable for which the mean (or average), **Categorical** — a variable consisting of a set of levels, **Unary** — a variable that has the same value for every observation in the data set**, Binary** — a variable that has only two possible levels, **Nominal** — a variable that has more than two levels, **Ordinal** — a variable that has more than two levels.

I. INTRODUCTION

According to Hand (1998), it is important for statisticians to become involved in data mining because statistical methods provide the glue that holds the process together. Because of its origins, data mining is a more practically oriented discipline than statistics. Therefore, it emphasizes topics that are not currently the focus of statistical research:

• Data mining deals with heterogeneous data, sometimes with complex internal structures (such as images, video, text, and signals).

• Data mining assumes that the raw data set is not of sufficient quality to apply a selected Statistical methodology directly. Instead, the data set must be prepared with appropriate Preprocessing techniques. That preparation can have as much or even more influence on the quality of the final results than the selected technique.

• Data mining uses flexible predictive techniques that are often based on strong algorithmic Foundations but have weaker formal statistical justification (such as artificial neural Networks and decision rules).

• Data mining often uses hidden (intermediate) variables as tools to perform a step-by-step compression of raw input data and presentation in more abstract form, which helps in building models for different tasks.

• Data mining attempts to find not only general, global models based on a data set but also local patterns (local models) in large data spaces, which is especially useful when the amount of data and the number of dimensions is so large that finding general models is cost prohibitive

• Data mining has a strong emphasis on an algorithmic approach and pays attention to issues of scalability (i.e., whether the approach will work with reasonable efficiency in large data sets). Approaches that are not practical are discouraged.

Data mining concentrates on data management and optimization of data searches (with a focus on problems of preprocessing, data cleaning, algorithms, and data structures). Statistics is more oriented toward formalisms for final model representation and score function formalization in the data space to perform inference (with a focus on problems of models and principles of statistical inference). At the same time, data miners have focused on estimation and have generally ignored inferential models. Table 1.1 briefly compares statistics and data mining in terms of specific problems and tools.

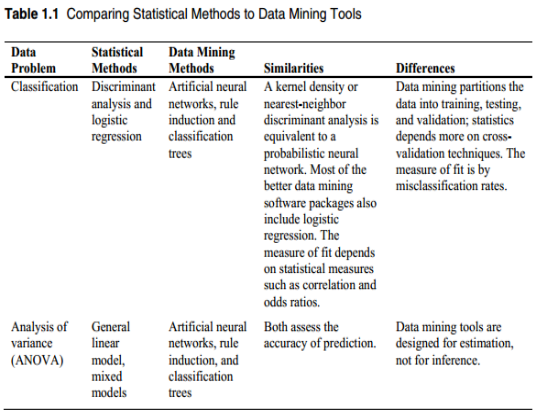


Table 1.1 Comparing Statistical Methods to Data Mining Tools

Data Problem	Statistical Methods	Data Mining Methods	Similarities	Differences
Classification	Discriminant analysis and logistic regression	Artificial neural networks, rule induction and classification trees	A kernel density or nearest-neighbor discriminant analysis is equivalent to a probabilistic neural network. Most of the better data mining software packages also include logistic regression. The measure of fit depends on statistical measures such as correlation and odds ratios.	Data mining partitions the data into training, testing, and validation; statistics depends more on cross-validation techniques. The measure of fit is by misclassification rates.
Analysis of variance (ANOVA)	General linear model, mixed models	Artificial neural networks, rule induction, and classification trees	Both assess the accuracy of prediction.	Data mining tools are designed for estimation, not for inference.

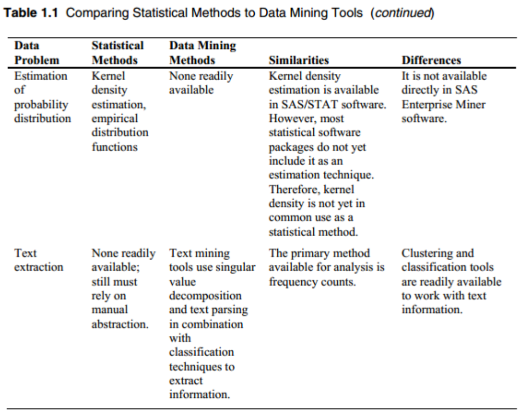


Table 1.1 Comparing Statistical Methods to Data Mining Tools (*continued*)

Data Problem	Statistical Methods	Data Mining Methods	Similarities	Differences
Estimation of probability distribution	Kernel density estimation, empirical distribution functions	None readily available	Kernel density estimation is available in SAS/STAT software. However, most statistical software packages do not yet include it as an estimation technique. Therefore, kernel density is not yet in common use as a statistical method.	It is not available directly in SAS Enterprise Miner software.
Text extraction	None readily available; still must rely on manual abstraction.	Text mining tools use singular value decomposition and text parsing in combination with classification techniques to extract information.	The primary method available for analysis is frequency counts.	Clustering and classification tools are readily available to work with text information.

Another important aspect of data analysis is the fitness and quality of data. In order to analyze the data statistically, it must first be in a format that we can analyze. According to Ferguson (1997), most organizations are drowning in data but starving for real information. As stated by Lambert (2002), many managers simply assume that the quality of the data they use is good. Unfortunately, poor quality appears to be the norm rather than the exception, and statisticians have largely ignored the issue. Poor quality data can cause more immediate harm and have other more indirect effects, according to Lambert (2002). Therefore, statisticians must consider the state of the data and the applicability of the statistical models. Working with complex data provides tremendous opportunities for statisticians who master the techniques of data mining.

Another major difference concerns the size of the data sets. In statistics, with the primary concern being inference, p-values and statistical significance are the primary measures of model effectiveness. However, data mining typically involves large data sets that are observational rather than random. The confidence width and effect size in such large samples decreases to 0 as the sample size increases. It is not unusual to have a regression model with every parameter statistically significant but with a correlation coefficient of almost 0 [1]. Therefore, other measures of model effectiveness are used. In data mining, the data sets are usually large enough to partition into three types: training, testing, and validation. The training data set is used to define the model, the testing data set is used in an iterative process to change the model if necessary to improve it, and the validation data set represents a final examination of the model. Depending upon the profit and loss requirements in the data, misclassification is used in supervised learning where there is a specific outcome variable [2].

Another issue with observational data generally is the problem of potential confounders. Although this problem can be dealt with using statistical linear models by including potential confounders in the model, in practice, the data sets collected are too small to include all confounders that should be considered. Because data mining tools can include

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

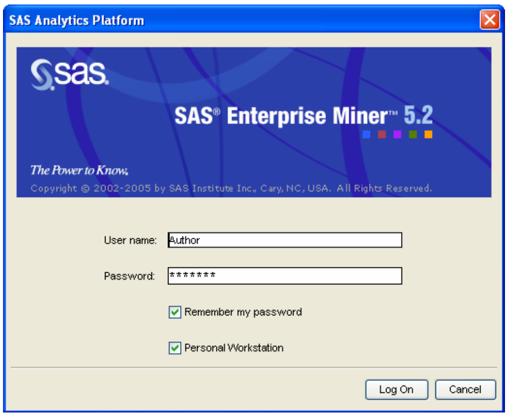
hundreds and sometimes thousands of variables simultaneously, potential confounders can and should always be considered in the data mining process.

II. LIREATURE SURVEY

SAS Enterprise Miner software streamlines the data mining process to create highly accurate predictive and descriptive models. The models are based on analysis of vast amounts of data from across an enterprise. Interactive statistical and visualization tools help you better search for trends and anomalies and help you focus on the model development process. This section discusses SAS Enterprise Miner 5.2 in detail.

Click **Log On** to start SAS Enterprise Miner software (Display 1.1).

Display 1.1 Initial SAS Enterprise Miner 5.2 Screen



Copyright © 2002-2005 by SAS Institute Inc., Cary, NC, USA. All Rights Reserved.

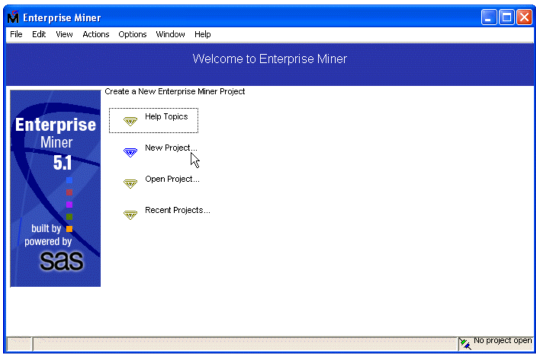
User name: Author

Password: *******

☑ Remember my password

☑ Personal Workstation

Display 1.2 Project Screen



After you have selected **New Project**, the window shown in Display 1.3 prompts you for a project name.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

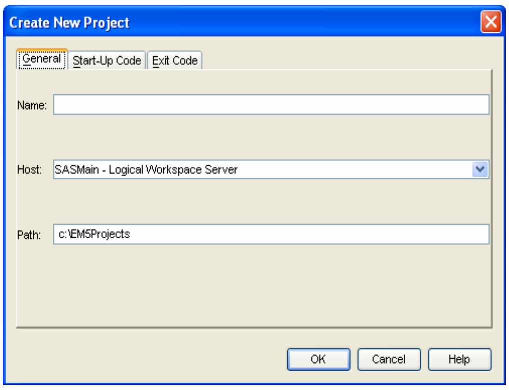
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Specify a name for the project in the **Path** field.

Display 1.3 Creating a New Project



Name:

Host: SASMain - Logical Workspace Server

Path: c:\EM5Projects

There are tabs for start-up code and for exit code. **Start-up code** allows you to enter SAS code that runs as soon as the project is open; **Exit Code** runs every time the project is closed [3]. The **Start-Up Code** tab is generally used to define a LIBNAME statement to inform SAS Enterprise Miner where all the project data are located [5]. Enter the following code here:

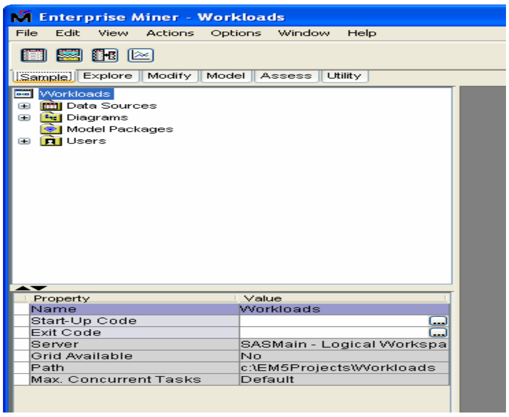
libname project 'C:\project_directory';

Once this code runs, all data sets created in the project are stored in the C:\project_directory folder. If desired, you can assign the Sasuser library name that is the default in Base SAS software [4]. The purpose of creating a separate library name for each project is to organize the data files. Only directories defined through the LIBNAME statement are recognized in the project [6]. Any data set you use must be located within the directory defined by the LIBNAME statement. Multiple LIBNAME statements can be used in the start-up code. However, at least one LIBNAME statement is required.

Once you have created a new project, the Enterprise Miner—Workloads window appears (see Display 1.4). The menus on the left change based on the choices you make.

.

Display 1.4 Initial SAS Enterprise Miner Project Screen



Property	Value
Name	Workloads
Start-Up Code	
Exit Code	
Server	SASMain - Logical Workspa
Grid Available	No
Path	c:\EM5Projects\Workloads
Max. Concurrent Tasks	Default

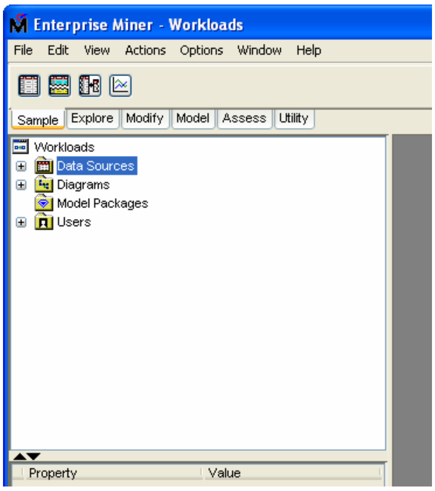
You can also access the start-up code directly through the project name menu, which is displayed by clicking the project name, to create a library name. Just click the … button next to **Start-Up Code** in the menu on the left [7].

All analyses begin with a data set. SAS Enterprise Miner software is used primarily to investigate large, complex data sets with tens of thousands to millions

of records and hundreds to thousands of variables. Data mining finds patterns and relationships in the data and determines whether the discovered patterns are valid [8].

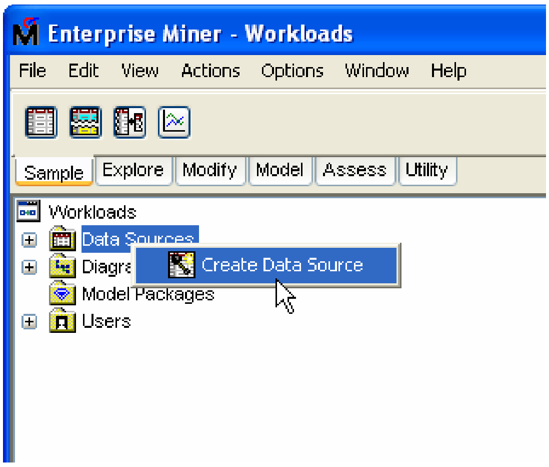
The next step is to access a data set. First, right-click **Data Sources**.

Display 1.5 Accessing Data Sets in a SAS Enterprise Miner Project


Enterprise Miner - Workloads
File Edit View Actions Options Window Help
Sample Explore Modify Model Assess Utility
Workloads
Data Sources
Diagrams
Model Packages
Users
Property Value

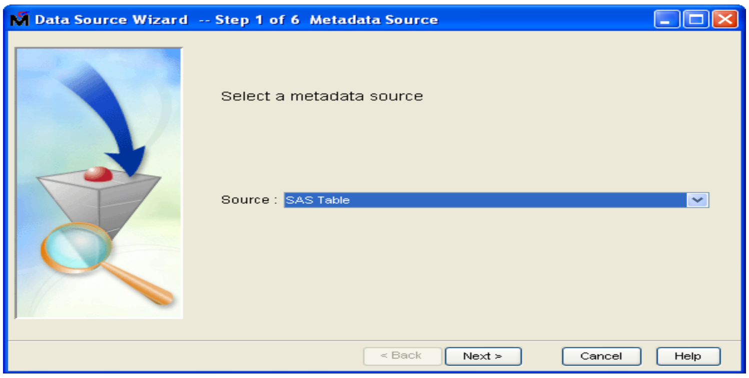
Click **Create Data Source** (Display 1.6).

Display 1.6 Creating a Data Source


Enterprise Miner - Workloads
File Edit View Actions Options Window Help
Sample Explore Modify Model Assess Utility
Workloads
Data Sources
Diagra
Model Packages
Users
Create Data Source

Next, a series of windows leads you through all the required steps to access a new data set in SAS Enterprise Miner 5.2.

Display 1.7 Creating a Data Source with the Data Source Wizard

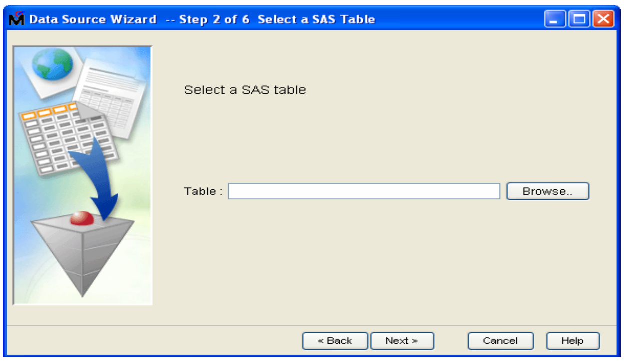


Select a metadata source

Source : SAS Table

The next step is to define the SAS data table. Click **Next** to accept the default (Display 1.8).

Display 1.8 Finding the Data File

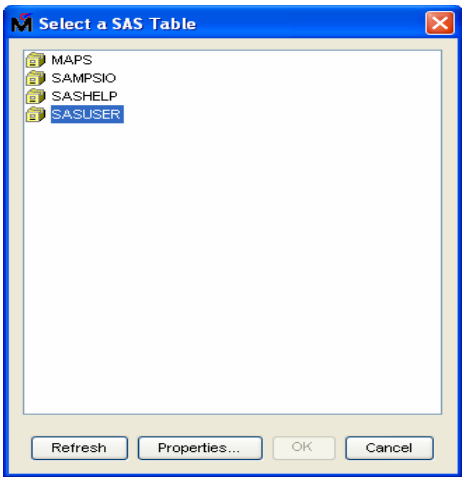


Select a SAS table

Table :

Click **Browse** to find data files contained within the defined libraries (Display 1.9).

Display 1.9 Defined Libraries



MAPS
SAMPSIO
SASHELP
SASUSER

If the libraries are not visible, refresh the screen. Select the library name where the data set is located and then click **OK**. Select the file, click **OK**, and then click **Next [9]**.

In Display 1.10, the general contents of the data file, including the number of variables and the number of observations, are listed. Display 1.10 also indicates when the data set was created and modified. Click **Next**.

Display 1.10 Data Properties

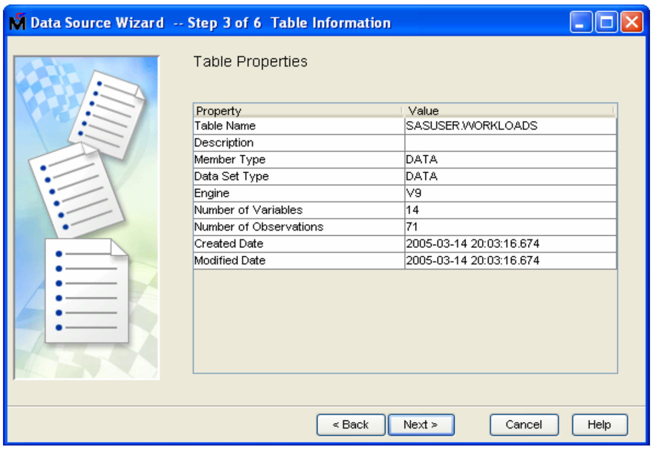


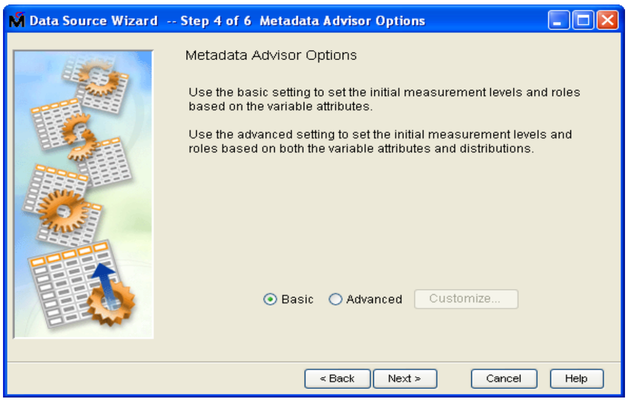
Table Properties

Property	Value
Table Name	SASUSER.WORKLOADS
Description	
Member Type	DATA
Data Set Type	DATA
Engine	V9
Number of Variables	14
Number of Observations	71
Created Date	2005-03-14 20:03:16.674
Modified Date	2005-03-14 20:03:16.674

Display 1.11 prompts you to choose basic or advanced information about the data set.

You should routinely choose **Advanced**, even though the default is **Basic**.

Display 1.11 Choice of Information Type



Metadata Advisor Options

Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.

● Basic ○ Advanced

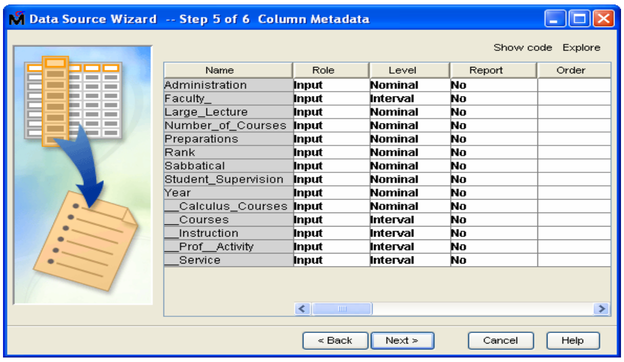
The variables in the data set are generally divided into four categories: identification, input, target, and rejected. A *rejected variable* is not used in any data mining analysis. An *input* (also called an *independent variable*) is used in the various models and exploratory tools in the software. Often, input variables are used to predict a *target value* (also called a *dependent value*). An *identification variable* is used to label a particular observation in the data set. While they are not used to predict outcomes, identification variables are used to link different observations to the same ID. Some data mining techniques require a target variable while others need only input variables [10].

There are other categories of variables as well that are used for more specialized analyses. In particular, a *time ID variable* is used in place of the more standard *ID variable* when the data are tracked longitudinally. A *text variable* is used with SAS Text Miner software [11]. A raw data set is used to perform the initial analyses, unless the data are longitudinal, in which case the data set is identified as transactional.

Many of the procedures used in SAS Enterprise Miner 5.2 are also available in SAS/STAT software. However, the procedures have different names, and sometimes perform slightly different functions. For example, the regression node in SAS Enterprise Miner performs linear or logistic regression, depending upon whether the target variable is continuous or discrete. The procedure itself determines the method to use. Similarly, there is a node for clustering. Other procedures, such as PROC NEURAL, are unique to SAS Enterprise Miner software [12]. Methods that are available in SAS/STAT software that are not readily available in SAS Enterprise Miner can be added through a code node, which allows you to copy and paste code as needed into the SAS Enterprise Miner window. The ability to integrate SAS Enterprise Miner with SAS/STAT software and other SAS components is one of the great strengths of SAS Enterprise
Miner software.

The sample data set shown in Display 1.12 identifies the role of all the variables as **Input**. However, not ll the variables are listed. Use the scroll bar to access the remaining variables [13].

Display 1.12 Variable Properties

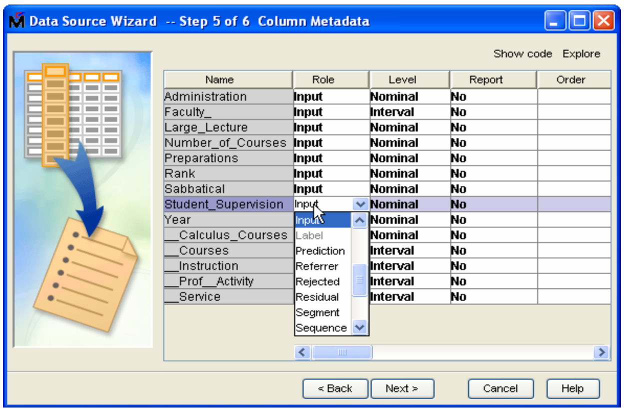


Name	Role	Level	Report	Order
Administration	Input	Nominal	No	
Faculty_	Input	Interval	No	
Large_Lecture	Input	Nominal	No	
Number_of_Courses	Input	Nominal	No	
Preparations	Input	Nominal	No	
Rank	Input	Nominal	No	
Sabbatical	Input	Nominal	No	
Student_Supervision	Input	Nominal	No	
Year	Input	Nominal	No	
__Calculus_Courses	Input	Nominal	No	
__Courses	Input	Interval	No	
__Instruction	Input	Interval	No	
__Prof__Activity	Input	Interval	No	
__Service	Input	Interval	No	

Initially, SAS Enterprise Miner 5.2 lists default role values. You can change these roles, as shown in Display 1.13.

For variables highlighted in the **Role** column, options are displayed. You can click on the role to be assigned to the variable [14]. In this example data set, only some of the variables will be examined; others will be rejected because they are not needed.

Display 1.13 Changing Variable Roles and Levels

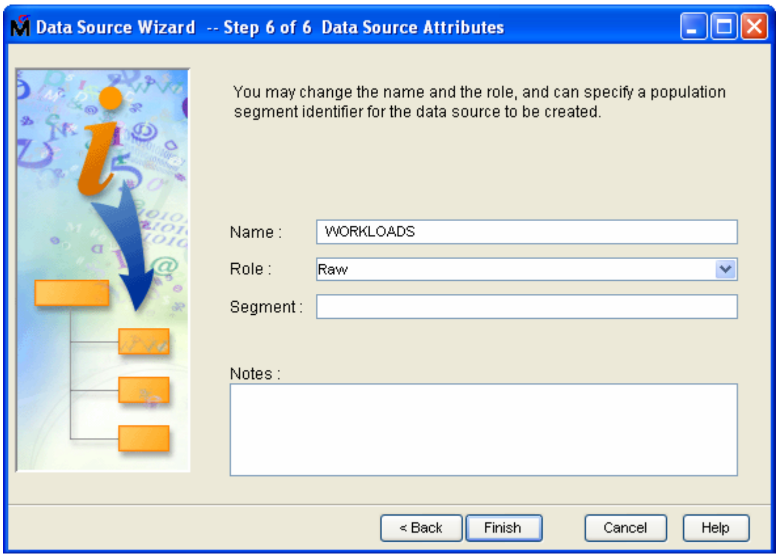


Name	Role	Level	Report	Order
Administration	Input	Nominal	No	
Faculty_	Input	Interval	No	
Large_Lecture	Input	Nominal	No	
Number_of_Courses	Input	Nominal	No	
Preparations	Input	Nominal	No	
Rank	Input	Nominal	No	
Sabbatical	Input	Nominal	No	
Student_Supervision	Input	Nominal	No	
Year		Nominal	No	
__Calculus_Courses		Nominal	No	
__Courses		Interval	No	
__Instruction		Interval	No	
__Prof__Activity		Interval	No	
__Service		Interval	No	

Other terms and roles can be used. However, this book focuses on the most basic and introduces others as needed.

Similarly, the level can be changed. Most of the levels in this data set are either **Nominal** or **Interval**. Not all of the data fit those levels. For example, the Number_of_Courses variable will not be **Nominal** as the variable name itself indicates that it is a number. Therefore, it will be changed to **Interval**. Display 1.14 shows the overall attributes and the role of the data file itself.

Display 1.14 Data File Attributes



You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : WORKLOADS

Role : Raw

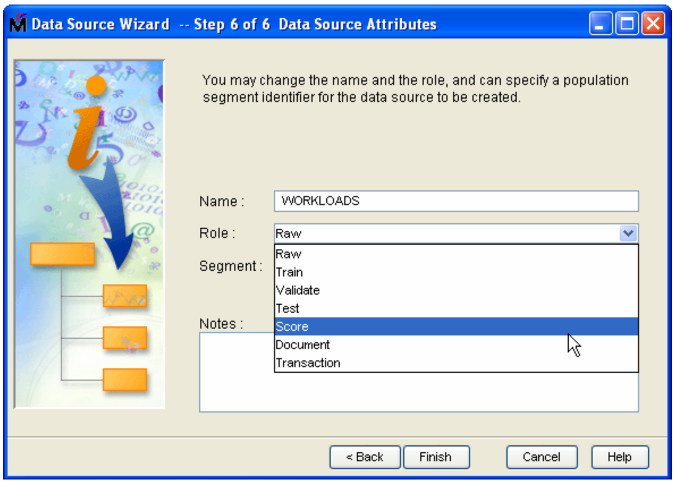
Segment :

Notes :

There are several roles for the data file, as shown in Display 1.15. They are explained in later chapters. For the basic analysis, the assigned role should be the default value of **Raw**.

Other possible data set roles include **Transaction**, which is used if there is a time component to the data.

Display 1.15 Possible Data File Attributes



You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : WORKLOADS

Role : Raw

Segment :

Notes :

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

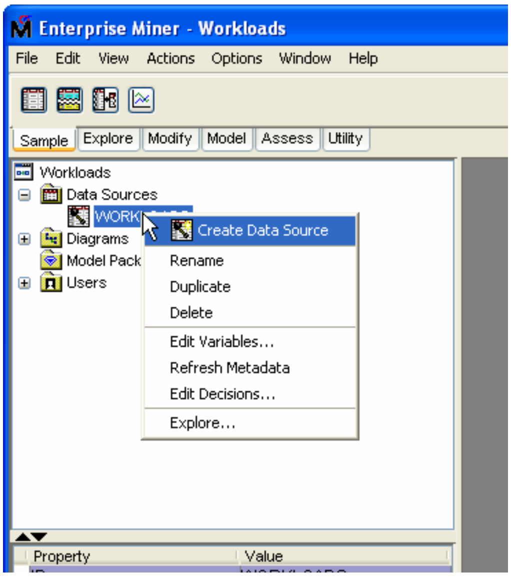
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Others include **Train**, **Validate**, **Test**, **Score**, and **Document**. **Score** is used when fresh data are introduced into a model. The remaining values will not be used in this example.

It is sometimes necessary to change either a variable role or a level in the data set by selecting **Edit Variables**, which is found by right-clicking the data set name in the drop-down list (Display 1.16).

Display 1.16 Changing Variable Roles and Levels



This option returns you to the window shown in Display 1.12.

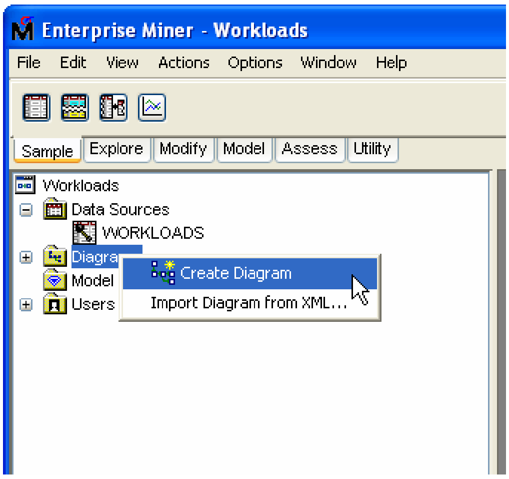
Exploring the Data Set

To perform data mining, first you need to understand the data. Therefore, you must explore the data, draw some graphs of the data, and perform basic data summaries. These summaries include some tables. Some of the exploration is relatively straightforward and can be performed using a number of SAS components, including SAS/INSIGHT software, SAS/STAT software, and SAS/GRAPH software. Some of the basics are available in SAS Enterprise Miner 5.2, and are shown here [15].

There are a number of ways to explore a data set in SAS Enterprise Miner 5.2. The first is to click **Explore** (shown in the upper right-hand corner of Display 1.12). A second way is to select **Explore** from the menu shown in Display 1.16. A third way is to click the Stat Explore icon located on the **Explore** tab. Each of these options leads to the same exploration. However, in order to use the Stat Explore icon, a diagram should be defined. The other two options do not require a diagram.

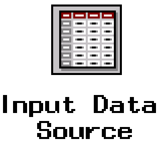
Once the data set is defined in the software, a diagram can be defined. The first step is to right click the Diagrams icon (Display 1.17).

Display 1.17 Defining a Diagram



You are then prompted to name the diagram (Display 1.18). Once you provide the name, the window on the right-hand side of the screen becomes white. The SAS Enterprise Miner diagram is constructed within that white frame.

Overview of the Nodes

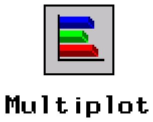


The Input Data Source node reads data sources and defines their attributes for later processing by Enterprise Miner. This node can perform various tasks:

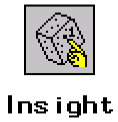
- Access SAS data sets and data marts. Data marts can be defined by using the SAS Data Warehouse Administrator, and they can be set up for Enterprise Miner by using the Enterprise Miner Warehouse Add-ins.
- Automatically create a metadata sample for each variable when you import a data set with the Input Data Source node. By default, Enterprise Miner obtains the metadata sample by taking a random sample of 2,000 observations from the data set that is identified in the Input Data Source. Optionally, you can request larger samples. If the data is smaller than 2,000 observations, the entire data set is used.
- Use the metadata sample to set initial values for the measurement level and the model role for each variable. You can change these values if you are not satisfied with the automatic selections that are made by the node.
- Display summary statistics for interval and class variables.
- Define target profiles for each target in the input data set.

Note: This document uses the term *data sets* instead of *data tables*.

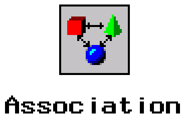
The Distribution Explorer node enables you to explore large volumes of data in multidimensional histograms. You can view the distribution of up to three variables at a time with this node. When the variable is binary, nominal, or ordinal, you can select specific values to exclude from the chart. To exclude extreme values for interval variables, you can set a range cutoff. The node also generates simple descriptive statistics for the interval variables [16].



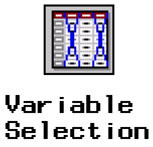
The Multiplot node enables you to explore large volumes of data graphically. Unlike the Insight or Distribution Explorer nodes, the Multiplot node automatically creates bar charts and scatter plots for the input and target variables without making several menu or window item selections. The code that is created by this node can be used to create graphs in a batch environment, whereas the Insight and Distribution Explorer nodes must be run interactively.



The Insight node enables you to open a SAS/INSIGHT session. SAS/INSIGHT software is an interactive tool for data exploration and analysis. With it, you explore samples of data through graphs and analyses that are linked across multiple windows. You can analyze univariate distributions, investigate multivariate distributions, and fit explanatory models by using generalized linear models.



The Association node enables you to identify association relationships within the data. For example, if a customer buys a loaf of bread, how likely is the customer to buy a gallon of milk as well? The node also enables you to perform sequence discovery if a time-stamp variable (a sequence variable) is present in the data set.



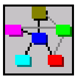
The Variable Selection node enables you to evaluate the importance of input variables in predicting or classifying the target variable. To select the important inputs, the node uses either an R-square or a Chi-square selection (tree-based) criterion. The R-square criterion enables you to remove variables that have large percentages of missing values, remove class variables that are based on the number of unique values, and remove variables in hierarchies. Variables can be hierarchical because of levels of generalization (Zipcode generalizes to State, which generalizes to Region) or because of formulation (variable A and variable B may have interaction A*B). The variables that are not related to the target are set to a status of rejected [6]. Although rejected variables are passed to subsequent nodes in the process flow diagram, these variables are not used as model inputs by more detailed modeling nodes, such as the Neural Network and Tree nodes. Certain variables of interest may be rejected by a variable selection technique, but you can force these variables into the model by reassigning the input model role to these variables in any modeling node.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015



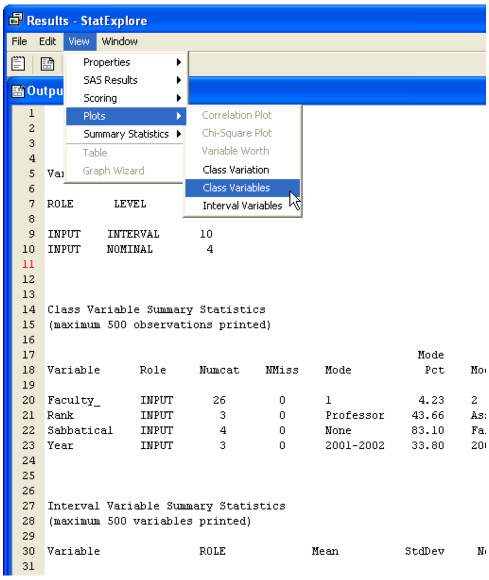
The Link Analysis node enables you to tranform data from different sources into a data model that can be graphed. The data model supports simple statistical measures, presents a simple interactive graph for basic analytical exploration, and generates cluster scores from raw data. The scores can be used for data reduction and segmentation

Analyzing a Sample Data Set

To explore the data well, you need to know about the domain. Therefore, it is essential to discuss the sample data set in more detail. This data set was chosen so that most readers would understand the basic domain of the data [5].

The data set contains the workload assignments for university faculty in one academic department for a three-year period. There are a total of 14 variables and 71 observations. There are a number of reasons to examine the data: to determine if employee resources are used efficiently and to determine whether there are longitudinal shifts in workload assignments that might impact overall productivity.

In terms of domain knowledge, it is important to understand why workloads are assigned in this fashion and why overall trends can be overlooked. Faculty members are responsible for publications, presentations, and grants at the end of the year. They are also responsible for teaching courses. As salaried employees, they have no required hours to fulfill on a daily or weekly basis. The workloads, then, are negotiated on an individual basis between each faculty member and administrative officials. Attempts to standardize workload requirements have not been entirely successful. Without standardization, trends are often missed because the data are not examined and summarized.



```
Var

ROLE       LEVEL

INPUT     INTERVAL       10
INPUT     NOMINAL         4



Class Variable Summary Statistics
(maximum 500 observations printed)

                                                          Mode
Variable      Role      Numcat    NMiss    Mode            Pct    Mo

Faculty_      INPUT       26        0      1               4.23   2
Rank          INPUT        3        0      Professor      43.66   As
Sabbatical    INPUT        4        0      None           83.10   Fa
Year          INPUT        3        0      2001-2002      33.80   20



Interval Variable Summary Statistics
(maximum 500 variables printed)

Variable               ROLE             Mean          StdDev     N
```

Presenting Additional SAS Enterprise Miner Icons

It is always possible to find patterns in the data. For example, flip a coin 10 times and suppose it comes up H, H, H, H, H, T, T, T, T, T. Without any attempt to verify the pattern of 5 heads followed by 5 tails, it is possible (although not valid) to conclude that every time a coin is flipped 10 times, the same pattern will occur. It is a conclusion that is easily contradicted if the coin is flipped several more times. Although this example seems somewhat absurd, other patterns seem to be accepted by many people with even less justification. Pattern recognition without validation is a reason that data mining as a method was often disparaged in statistical training [7].

Therefore, it is strongly recommended that you partition the data routinely into three data sets: training, validation, and testing. The training data set defines the model in SAS Enterprise Miner 5.2. The validation data set iteratively ensures that the developed model fits a fresh data set. Once the model is completed, the testing data set makes a final comparison. Because the division of a data set is so crucial to validation of the process, the software is set up so that splitting the data set into three components is almost automatic.

For a given target value, the accuracy of the final model is initially judged by the misclassification rate, where misclassification occurs when the predicted target value is not equal to the actual target value. There are additional diagnostics in the software that are discussed later.

Another difference between traditional statistics and data mining is that there are often many different models that can be used to investigate the data. Instead of choosing just one model to define a specific *p*-value, many different models are used and compared. Assessment methods have been developed to make these comparisons using the training, validation, and testing methodology.

Another important component of SAS Enterprise Miner 5.2 is the ability to score data. Scoring relates the predicted value to the actual value, and the closeness of one to the other can be examined using other statistical techniques [8]. Scoring is particularly important when examining the likelihood of a customer making a particular purchase and the amount of the purchase. In this case, scoring assigns a level of importance to a customer. For wireless phone service, churn (meaning that a customer decides to switch phone providers) is important; the provider wants to predict in advance those who are likely to churn in order to give those customers incentives to stay. How can a business predict the likelihood of churn to offer incentives to prevent it? Scoring provides a means of making such predictions.

A number of icons on the **Sample** and **Modify** tabs are useful in investigating data. The Partition icon (Display 1.28) is extremely important and should be used almost routinely, particularly with very large data sets. This icon divides the data into three sets: train, validate, and test [8]. For many of the models in the software, the training data set initially defines the model; the validation data set tests the model in an iterative process with the training set. The testing data set examines the overall accuracy of the model once it is complete.

III. conclusion

You should investigate the Workload data set using the SAS Enterprise Miner 5.2 nodes discussed in this chapter. A second data set is provided, called Student_Survey. The variables in the list correspond to the order of the questions asked in the following survey. The variable names correspond to the questions in the survey. However, all possible responses to question 1 are listed as separate variables with a binary response. Open-ended questions were not included in the data set. Question 19 is an interval response; all others are ordinal or nominal responses. This data set is used in subsequent chapters. Therefore, you should become familiar with it by using the StatExplore and MultiPlot icons to examine the data. There are many missing values in the data set, so imputation is required. Some errors require auditing of the data. Because most variables are similarly scaled, standardization is not required (unless the Hours variable is included in an analysis). There are approximately 190 observations, so it is reasonable to partition the data set.

REFERENCES

1. Ferguson, Mike. 1997. “Evaluating and Selecting Data Mining Tools.” *InfoDB* 11(2): 1–10.
2. Udayakumar R., Khanaa V., Saravanan T., "Analysis of polarization mode dispersion in fibers and its mitigation using an optical compensation technique", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4767-4771.
3. Hand, David J. 1998. “Data Mining Statistics and More?” *The American Statistician* 52(2):112–118.
4. Udayakumar R., Khanaa V., Saravanan T., "Chromatic dispersion compensation in optical fiber communication system and its simulation", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4762-4766.
5. Lambert, Bob. 2002. “Data Warehousing Fundamentals: What You Need to Know to Succeed.”*DM Review*. Intelligence Platform: Data Administration Guide, available from

6. Intelligence Platform: Desktop Application Administration Guide, available from SAS® Deployment Wizard User's Guide. SAS Institute Inc.: Cary, NC. Available at support.sas.com/documentation/installcenter/en/ikdeploywizug/62130/PDF/default/user.pdf.
7. Udayakumar, R., Khanaa, V., Saravanan, T., "Synthesis and structural characterization of thin films of sno2 prepared by spray pyrolysis technique", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp.4754-4757.
8. SAS Institute Inc. 2010 "Standard Deployment Plans: Enterprise Guide, two machines." SAS Institute Inc.: Cary, NC.
9. Uma Mageswaran, S., Guna Sekhar, N.O., "Reactive power contribution of multiple STATCOM using particle swarm optimization", International Journal of Engineering and Technology, ISSN : 1793-8236, 5(1) (2013) pp. 122-126.
10. Harrison, A.: Design for service harmonizing product design with a services strategy.
11. Wills, G., Fowler, D., Sleeman, D., Crowder, R., Kampa, S., Carr, L., Knott, D.: Issues in moving to a semantic web for a large corporation
12. Wong, S.C., Crowder, R.M., Wills, G.B., Shadbolt, N.R.: Knowledge engineering - from front-line support to preliminary design. In Brailsford, D.F., ed.:
13. Gao, Y., Zeid, I., Bardasz, T.: Characteristics of an effective design plan system to support reuse in case-based mechanical design.
14. Vidyalakshmi K., Kamalakannan P., Viswanathan S., Ramaswamy S., "Antinociceptive effect of certain dihydroxy flavones in mice", Pharmacology Biochemistry and Behavior, ISSN : 0091-3057, 96(1) (2010) pp. 1-6.
15. Khadilkar, D.V., Stauffer, L.A.: An experimental evaluation of design information reuse during conceptual design..
16. Jagtap, S., Johnson, A., Aurisicchio, M., Wallace, K.: Pilot empirical study: Interviews with product designers and service engineers.
17. R.Karthikeyan, Dr.S.R.Suresh, An Approach for Real Time Testing Reliability& Usability Testing Process, International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Online): 2320 – 9801,pp 644-651, Vol. 1, Issue 3, May 201
18. P.Ramya, dr.nalini, effective navigation queryresults based on biomedicaldatabase, International Journal of Innovative Research in Computer and Communication Engineering, ISSN: 2249-0183,pp 36-39, volume 1 Issue 3 No4- Dec 2011
19. P.Kavitha, Authentic Learning Activities withPedagogical Stylistics – Enhancing InE-Learning Websites, International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801,pp 1211-1215, Volume 1, Issue 6, August 2013
20. N.Priya, VANET Based Adaptive Traffic SignalControl, International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801,pp 1201-1208, Vol. 1, Issue 6, August 2014
21. N.Sakthi Priya, Cervical Cancer Screening and ClassificationUsing Acoustic Shadowing, International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801,pp 1676-1679, Volume 1, Issue 8, October 2013