Finding the Solution to Data Mining

Exploring the Features and Components of Enterprise Miner™, Release 4.1 from SAS®

The Power to Know™

Table of Contents

The Primary Content Providers for *Finding the Solution to Data Mining* were Ron Statt and Wayne Thompson in SAS' Cary, North Carolina, offices.

Introduction

Data mining is a process, not just a series of statistical analyses. Implementing a successful data mining process requires a solution that gives users:

- Advanced, yet easy-to-use, statistical analyses and reporting techniques.
- A guiding, yet flexible, methodology.
- Client/server enablement.

Enterprise Miner software from SAS *is* that solution. It synthesizes the world-renowned statistical analysis and reporting systems of SAS software with an easy-to-use graphical user interface (GUI) that can be understood and used by business analysts as well as quantitative analysts.

This paper presents the three main features of Enterprise Miner software: a single GUI, a process for data mining and client/server capabilities.

Enterprise Miner Software

Enterprise Miner provides powerful data mining capabilities that meet the needs of a wide variety of users, including information technology staff, business analysts and quantitative analysts. With Enterprise Miner, users can:

- Identify the most profitable customers and the underlying reasons for their loyalty.
- Determine why customers leave to use competitors.
- Analyze clickstream data to improve e-commerce strategies.
- Increase customer profitability and reduce risk exposure through more accurate credit scoring.
- Determine the combination of products customers are likely to purchase and when.
- Set more profitable rates for insurance premiums.
- Save on downtime by applying predictive maintenance to manufacturing sites.
- Detect and deter fraudulent behavior.

Enterprise Miner provides the most complete set of data mining tools for data preparation and visualization, predictive modeling, clustering, association discovery, model management, model assessment and reporting. A key advantage of using Enterprise Miner is that it generates complete scoring code for all stages of model development. Scoring code can be applied immediately to new data on any SAS system, which results in significant time savings. Scoring

code is also available in the C and Java languages to enable users to deploy results outside of the SAS environment.

Because it's a software solution for business from SAS, Enterprise Miner can integrate with other SAS software, including the award-winning SAS/Warehouse Administrator software, the SAS solution for online analytical processing (OLAP), and SAS/IntrNet software.

The Graphical User Interface

Enterprise Miner uses a single GUI to provide the functionality that users need to uncover valuable information hidden in their data. With one point-and-click interface, users can perform the entire data mining process, from sampling data, through exploration and modification, to modeling, assessment, and scoring of new data for use in making business decisions.

The GUI is designed with two groups of users in mind. Business analysts, who might have limited statistical expertise, can use the GUI to quickly and easily navigate through the data mining process. Quantitative analysts, who might want to explore the details, can use the GUI to access and fine-tune the underlying analytical processes.

The GUI Components

The primary components of the GUI, as shown in Figure 1, are:

- **Project Navigator.** Use to manage projects and diagrams; add tools, such as analysis nodes to the diagram workspace; and view HTML reports that are created by the Reporter node. The Tools sub-tab of the Project Navigator displays the data mining nodes that are available for constructing process flow diagrams (PFDs). An example is shown in Figure 2.

- **Diagram Workspace.** Use for building, editing and running PFDs.

- **Tools Palette.** Contains a subset of the Enterprise Miner tools that are commonly used to build PFDs in the diagram workspace. Users can add to and delete from the Tools palette.

- **Nodes.** A collection of tools that enable users to perform tasks in the data mining process, such as data access, data analysis and reporting. Nodes have a uniform look-and-feel that makes them easy to use.

- **Message Panel**, **Progress Indicator** and **Connection Status Indicator** — use to display information and task completion estimates in response to user actions. These components appear across the bottom of the workspace window.

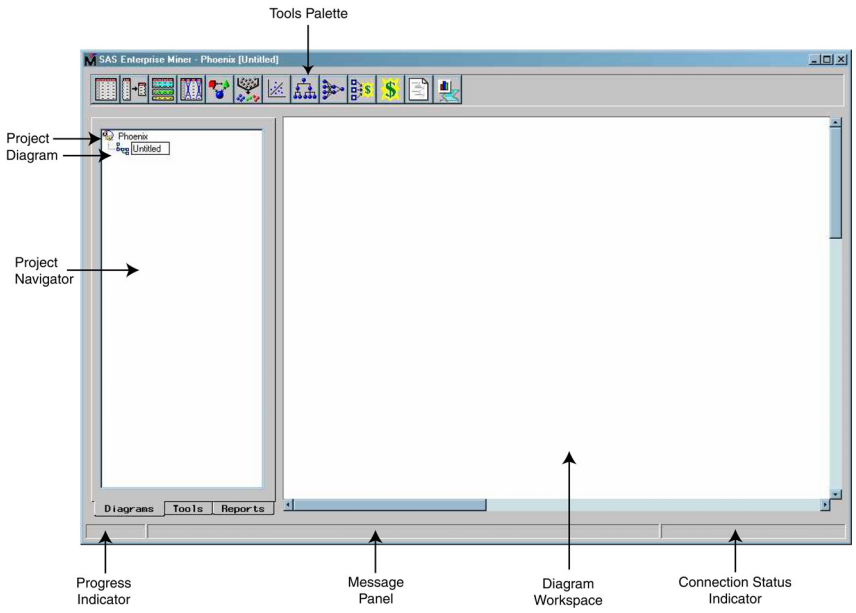

Figure 1: Enterprise Miner GUI

With these easy-to-use tools, users can map an entire data mining project, execute individual functions and modify PFDs.

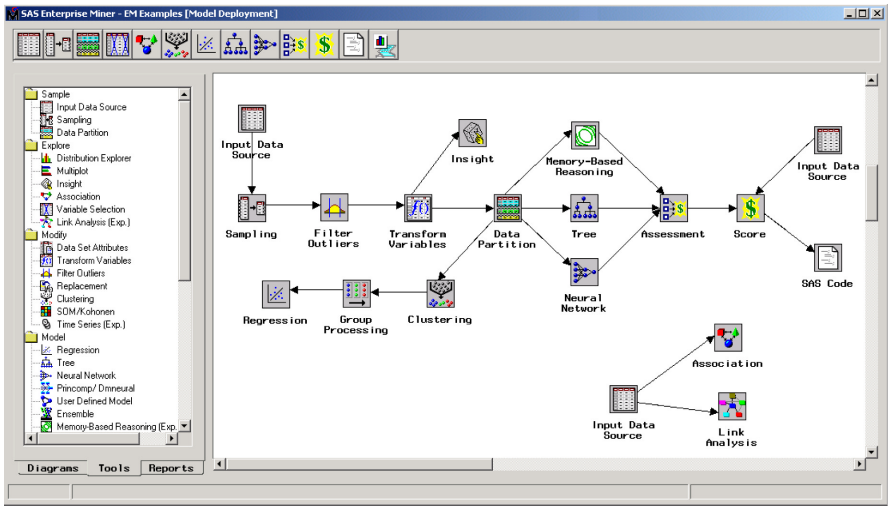

Figure 2: Tools sub-tab of the Project Navigator and example of Process Flow Diagram

The Enterprise Miner Process for Data Mining

The Enterprise Miner GUI is effective because it enables business analysts and quantitative analysts to organize data mining projects into a logical framework. The visible representation of this framework is a PFD, which graphically illustrates the tasks performed to complete a data mining project. The PFD also eliminates the need for manual coding and dramatically lessens the time needed to develop models.

In addition, a larger, more general framework for staging data mining projects exists. SAS defines this framework as SEMMA, that is, a proven methodology that enables users to sample, explore, modify, model and assess data through an easy-to-use GUI.

SEMMA provides users with a powerful and comprehensive method in which individual data mining projects can be developed and maintained. For more information about SEMMA, see the SAS white paper “Data Mining with the SAS System: From Data to Business Advantage.”

The GUI nodes are primarily organized according to the SEMMA methodology. However, users are not required to follow the SEMMA methodology exactly. The relationship between the nodes and the methodology is flexible, which allows for the repetition or the reordering of SEMMA tasks.

The Sampling Nodes

In data-intensive tasks such as data mining, intelligent sampling can greatly reduce the time required for processing. If it can be ensured that the sample data is sufficiently representative of the whole, patterns that appear in the entire database will also appear in the sample. Although Enterprise Miner does not require the use of sample data, sampling nodes can be used to develop representative sample data.

Input Data Source Node

The Input Data Source node enables users to access SAS data sets and views for use as input for data mining projects. Because Enterprise Miner can work in conjunction with SAS/ACCESS engines, users can read more than 50 different file formats, such as DB2 or Oracle tables.

With the Input Data Source node, a metadata sample is downloaded to the client, and crucial information about the data is automatically determined. The metadata (for example, data set roles, model roles, measurement levels and summary statistics) is used in the process flow to maximize ease-of-use and performance. As with all other nodes, the user can manually override any of the default or automatic settings.

Sampling Node

The Sampling node enables users to extract a sample of the input data source. Sampling is recommended for extremely large data sets, because it can significantly decrease model training time and produce more reliable models. When the target event is rare, it is often necessary to create a biased sample to produce useful models. The Sampling node supports simple random

sampling, sampling every *n*th observation, sampling the first *n* observations and more intelligent sampling techniques, such as stratified sampling and cluster sampling.

Data Partition Node

The Data Partition node enables users to partition the input data source or sample into data sets for training, validation and testing purposes. Partitioning the input data helps to speed preliminary model development. Additionally, partitioning provides mutually exclusive data set(s) for cross validation and model assessment. The partitioning can be based on simple, stratified random, or user-defined sampling. The user can specify the proportion of sampled observations to write to each partitioned data set.

The Exploration Nodes

The Exploration nodes in Enterprise Miner enable users to dynamically and iteratively explore and modify their data.

Distribution Explorer Node

The Distribution Explorer node is a fully interactive, advanced visualization tool that enables users to explore large volumes of data using multi-dimensional histograms. It can be used to visually identify patterns and trends in the data that reveal extreme values.

This node can also be used to explore data distributions created during other phases of the SEMMA process. For example, a user can assess model results by creating a chart of the predicted values versus the actual values from a modeling node.

Multiplot Node

The Multiplot node is a visualization tool that enables users to explore large volumes of data graphically. Unlike the Distribution Explorer or Insight nodes, the Multiplot node automatically

creates bar charts and scatter plots for the input and target variables in the data set. The code created by the Multiplot node can be used to create graphs in a batch environment.

With this node, users can annotate interval input by interval target scatter plots with a regression line and 90% confidence intervals for the mean. Users can also create stacked bar charts that illustrate the distribution of the target events throughout the range of an input variable.

Insight Node

The Insight node enables users to interactively explore and analyze data by means of multiple graphs and analyses that are linked across multiple windows. For example, users can analyze univariate distributions, investigate multivariate distributions, create scatter-and-box plots, display mosaic charts and examine correlations. In addition, users can fit explanatory models using analysis of variance (ANOVA), regression and the generalized linear model (GLM).

Association Node

The Association node enables users to identify association relationships within the data by performing either association or sequence discovery. Association discovery enables users to identify items that occur together in a specific event or record. The technique is also known as *market basket analysis*. Rules of form are discovered based on frequency counts of the number of times relevant items occur alone and in combination in a database.

Sequence discovery, unlike association discovery, takes into account the temporal ordering of the relationships (time sequence) among items. For example, sequence discovery rules might indicate relationships such as, “Of those customers who currently hold an equity index fund in their portfolio, 15% will open an international fund in the next year.”

Figure 3: Grid plot of association rules. The support for each rule determines the color of the symbol and the confidence interval establishes the shape of the symbol.

Variable Selection Node

The Variable Selection node provides an algorithm to evaluate the importance of input variables in predicting or classifying the target variable. To preselect the important inputs quickly, the Variable Selection node uses either an R-square or a tree-based Chi-square selection criteria. Within these criterion, there are options that enable users to exclude variables unrelated to the target, exclude variables in hierarchies, exclude variables that have large percentages of missing values and that exclude class variables based on the number of unique values. The remaining information-rich inputs can be passed to one of the modeling nodes, such as the Neural Network node, for more detailed evaluation.

All criteria options are used by default for variable selection. They work together to determine which variables to eliminate. In addition, the Variable Selection node automatically applies intelligent grouping techniques to the input variables to allow for a nonlinear relationship to the target. For example, interval inputs are grouped into 16 classes and evaluated using an analysis of variance approach. Class variables can be grouped so that classes with similar relationships to the target are combined into a single class.

Link Analysis Node

Link Analysis is a technique that reveals the structure and content of a body of information by representing it as a set of interconnected, linked objects. Some applications include forms of fraud detection, criminal network conspiracies, telephone network analyses, Web site structure and usage, database visualization and social network analyses.

The Link Analysis node transforms data from various 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 that can be used for data reduction and segmentation.

The Link Analysis node supports several training methods. All training methods result in two data sets: NODES and LINKS.

NODES. This data set contains a row for every node in a link graph. Each row contains a unique ID variable.

LINKS. This data set contains a row for every link in a link graph. Each row must contain ID1 and ID2 variables that have valid values taken from the ID variable of the NODES data set.

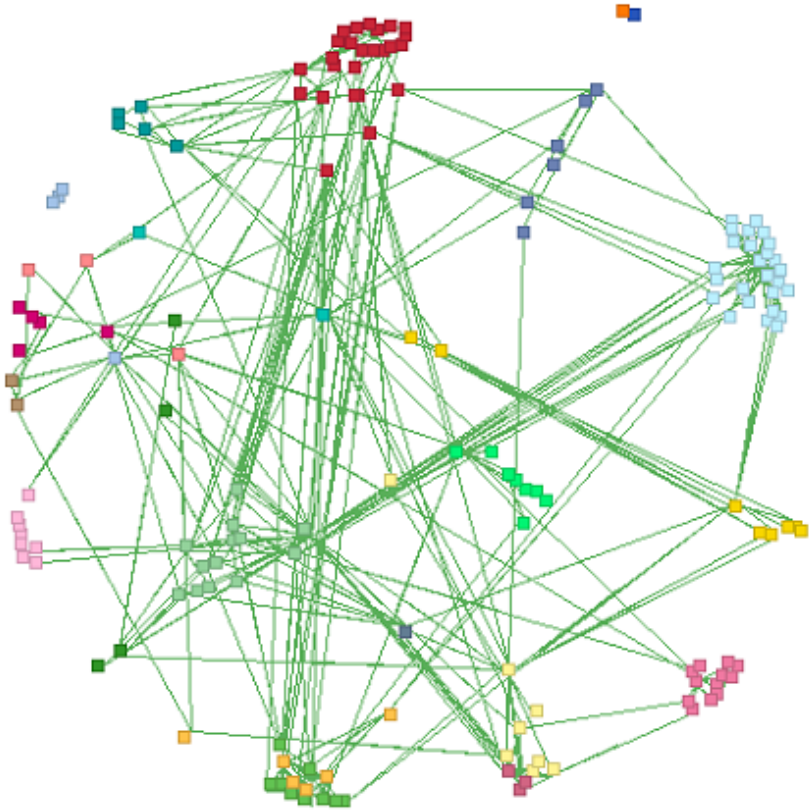

Figure 4: Clustered structure of Web pages resulting from Link Analysis

The Modifying Nodes

The modifying nodes enable users to adjust their data using their business and statistical rules.

Data Set Attributes Node

The Data Set Attributes node enables users to modify data set attributes (such as data set names, descriptions and roles) at any point within the PFD. This node can also be used to modify the metadata sample that is associated with a data set, for example, to change the target variable of a model.

Transform Variables Node

The Transform Variables node enables users to create new variables that are transformations of existing variables in your data. Transformations can help improve the fit of a model to the data. For example, this node provides automatic transformation algorithms that can be used to stabilize variances, remove non-linearity and correct non-normality in variables.

Users can choose from a predefined set of transformation types, including log, square root, inverse, square, exponential, standardize, bucket, quantile, binning and more. Additionally, users can create their own transformation or modify an existing transformation.

Filter Outliers Node

The Filter Outliers node enables users to apply a filter to data to exclude observations, such as outliers or other observations that are not wanted for further data mining analysis. Filtering extreme values from the data tends to produce better models because the parameter estimates are more stable. The filters can be based on business rules, such as the exclusion of customers older than 100 years, or based on statistical rules, such as the exclusion of observations that are outside the range of *x* standard deviations.

With this node, users can apply automatic filters that eliminate:

- Rare values for classification variables with fewer than 25 unique values.
- Extreme values for interval variables.
- Missing values.

Replacement Node

The Replacement node enables users to fill in values for observations that have missing values. This node provides a broad range of imputation statistics, including: none, set value, default constant, mean, median, midrange, distribution based and the mode. The Replacement node also provides automatic imputation techniques that predict missing values from the non-missing information in the input data by using decision trees and the ability to customize the default replacement statistics by specifying replacement values for missing and non-missing data.

If the missing values for observations are not filled in, then Enterprise Miner discards those observations for modeling by the Variable Selection, Neural Network, and Regression nodes. Discarding incomplete observations often means discarding the useful information that is contained in the non-missing values. This might also bias the sample.

Clustering Node

The Clustering node performs observation clustering to identify data observations that are similar according to some analytical metric.

The clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. The observations are divided into clusters such that every observation belongs to only one cluster.

After clustering is performed, the characteristics of the clusters can be interactively profiled. The results rank the input variables by their importance for the segmentation. Additionally, the cluster identifier for each observation can be passed to other nodes for use as an input, id, group or target variable. For example, a user can form clusters based on different age groups that will be targeted. Then, predictive models can be built for each age group by passing the cluster variable as a group variable to a modeling node.

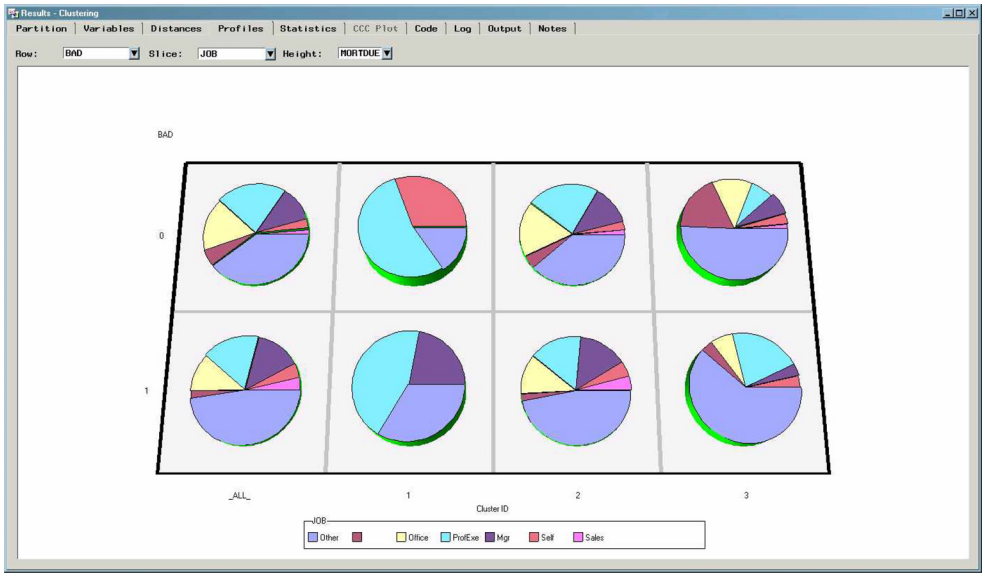

Figure 5: Cluster profile plot. A graphical representation of categorical and interval variables for each cluster. Height indicates interval variables. Slices and rows indicate categories by cluster and segment.

SOM/Kohonen Node

The SOM/Kohonen node performs unsupervised learning by using Kohonen vector quantization (VQ), Kohonen self-organizing maps (SOMs), or batch SOMs with Nadaraya-Watson or local-linear smoothing. Kohonen VQ is a clustering method, whereas SOMs are primarily dimension-reduction methods.

As with the Clustering node, after the network maps have been created, the characteristics of the clusters can be profiled graphically and cluster IDs can be passed on to subsequent nodes. Additionally, this node provides a report that indicates the importance of each variable.

Time Series Node

The Time Series node enables users to convert transactional data to time series data. Transactional data is time-stamped data that is collected over a period of time at no specific frequency. By contrast, time series data is time-stamped data that is collected over a period of time at a specific frequency (hours, days, weeks and more).

The Time Series node condenses the information that is contained in a transactional data set. This enables users to discover trends and seasonal variations that might not be visible in transactional data.

Users can specify a time unit and, optionally, a time range for their data. The Time Series node then transforms the data into time series format. As a result, users can view seasonal and trend plots of their data. More importantly, the data set can be used for further analysis, such as predictive modeling or clustering.

The Modeling Nodes

Enterprise Miner provides several modeling nodes to find good rules (models) for predicting the values of a target (dependent) variable from the values of the input (independent) variables. The methods used in these modeling nodes come from several different areas of research, including statistics, pattern recognition and machine learning.

The extent to how well a model fits the data is traditionally measured by a statistical criterion such as the average squared error or misclassification rate in the validation data. Alternatively, each modeling node can evaluate and select a candidate model by using a set of profit (or loss) functions, which can be in the form of a decision matrix. A decision matrix is created by specifying a profit (or loss) for each value of the target. When a model makes a prediction for an observation, each function is applied to the prediction, and the model selects the function that has the largest

profit (or the smallest loss). The selected function is called the decision, and the set of candidate functions is the set of decision alternatives. Therefore, an Enterprise Miner modeling node not only makes a prediction for an observation, it makes a decision about the observation, and estimates the profit or the loss. A profit function might become more realistic by subtracting a cost that depends on the individual observation. In this case, each model estimates the profit over the cost, or the return on investment, for each observation.

After a good model has been found, it can be applied to new data sets (scoring). All the predictive modeling nodes enable you to score the training, validation, test and scoring data sets in conjunction with training.

Regression Node

The Regression node enables users to fit both linear and logistic regression models. Based on the measurement level of the target, the node automatically applies the correct regression model to the input data. Linear regression attempts to predict the value of a continuous target as a linear function of one or more independent inputs. Logistic regression attempts to predict the probability that a binary or an ordinal target will acquire the event of interest as a function of one or more independent inputs.

The Regression node supports forward, backward and stepwise selection methods. The user can choose a selection criterion, such as minimize the average square error or maximize the classification rate in the validation data, to prevent overfitting of the training data. This node also includes a point-and-click "Interaction Builder" to assist you in creating higher-order modeling terms. Data sets that have a role of score are automatically scored when you train the model. In addition, Enterprise Miner enables users to build regression models in a batch environment.

The Regression node creates a comprehensive set of results, both in tabular and graphical format, which enables the user to easily assess the goodness-of-fit for the regression model.

Tree Node

A decision tree is a type of predictive model that partitions large amounts of data into segments that are called terminal nodes or leaves. The Tree node enables users to create decision trees that either:

- Classify or predict observations.
- Predict the appropriate decision when decision alternatives are specified.

As a result, decision trees produce a set of rules that can be applied to new data to generate predictions that can be used to drive business decisions. These rules can often be more easily interpreted from a business standpoint. Therefore, the rules themselves are more easily used in the business process.

The Tree node finds multi-way splits based on nominal, ordinal, and interval inputs. In addition to the default settings, users can choose the splitting criteria manually. The wide range of Tree node options also enables the user to mix algorithmic strategies advocated by Kass (1980) and by Breiman, et al. (1984). The Tree node extends the *p*-value adjustments of Kass and the retrospective pruning and misclassification costs of Breiman et al.

In addition, the Tree Node enables users to:

- Use prior probabilities and frequencies to train data.
- Base the criterion for evaluating a splitting rule on a statistical significance test.

The user has several tree-pruning options for deriving the optimal set of decision rules. By default, the trained tree is pruned based on performance using the validation data set. Additionally, the user can choose between keeping all or a specified number of leaves in the final model.

The Tree node supports both automatic and interactive mode. When the node is run in automatic mode, it ranks the input variables based on the strength of their contribution to the target. Interactive mode enables the exploration and evaluation of a large set of trees that is developed heuristically.

The Tree node produces extensive output that is organized in an interactive results browser. The user can view a summary table, a tree ring navigator, an assessment table and the tree diagram that graphically displays the decision rules.

Output from the Tree node can also be used outside of the SAS environment by using the Tree Results Viewer. This viewer is a Microsoft Windows application that provides novel visualization techniques to help find interesting leaves and problematic partitions. The Tree Results Viewer displays 14 linked tables and graphs in separate windows.

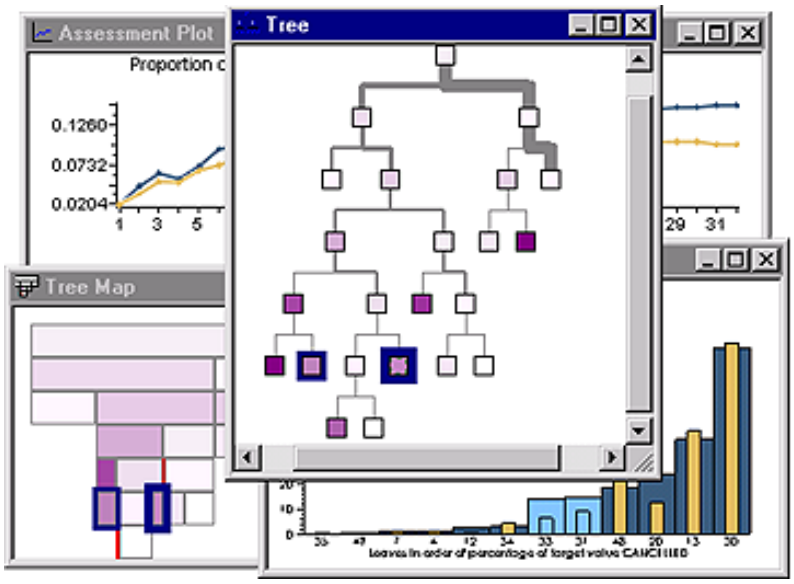

Figure 6: Tree Results Viewer. Selecting nodes, variables, or subtrees in a window automatically updates or selects corresponding elements in other windows.

Neural Network Node

An artificial neural network is a computer application that attempts to mimic the neurophysiology of the human brain in the sense that the network can find patterns in data from a representative data sample. More specifically, it is a class of flexible, nonlinear regression models; discriminant models; and data reduction models that are interconnected in a nonlinear dynamic system. By detecting complex nonlinear relationships in data, neural networks can help users make predictions about real-world problems.

The following neural network architectures are available in Enterprise Miner:

- Generalized linear model (GLM).
- Multi-layer perceptron (MLP), often the best for prediction problems.
- Radial basis function (RBF), often the best for clustering problems.
- Equal-width RBF.
- Normalized RBF.
- Normalized equal-width RBF.

The GUI of the Neural Network node enables novice and advanced users alike to make appropriate modeling selections. For the novice user, most of the required network parameters

are selected automatically based on the structure of the data and a few, easy user settings. The advanced user, however, can drill down and change all network parameters by using the GUI, which detects inadequate changes of interdependent parameters.

To avoid the tendency of neural networks to overfit the training data, the model performance is constantly assessed against the validation data, and the final model is selected based on one of several criteria that the user can select, such as the minimal validation error or maximum total profit. By using the interactive results browser, the user can assess the performance of the models at every iteration and change the final model selection manually, if they want to.

Output from the Neural Network node includes the following:

- Estimates data sets, for preliminary optimization and training.
- Output data sets, for training, validation, testing and scoring.
- Fit statistics data sets, for training, validation and testing.

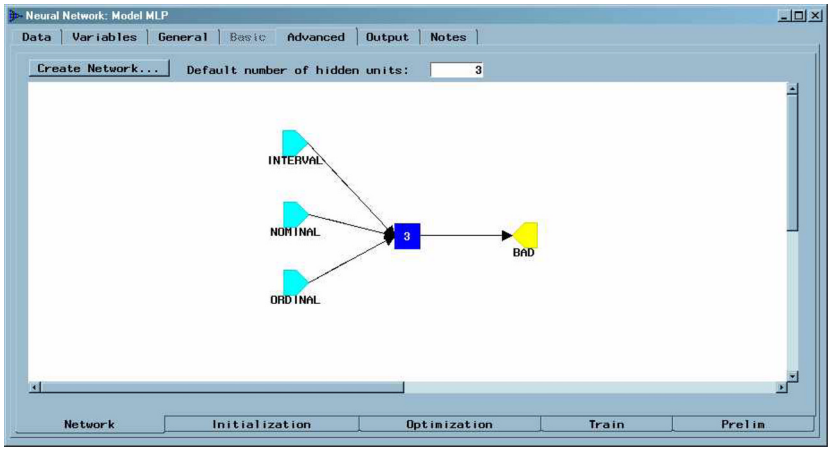

Figure 7: The neural network node provides several flexible and easy-to-define neural network architectures. For example, the user can insert hidden layers and specify the number of neurons, define skip layers, and choose from several activation and combination functions.

Princomp/Dmneural Node

The Princomp/Dmneural node enables users to fit an additive nonlinear model that uses bucketed principal components as inputs to predict a binary or an interval target variable. The node can also perform a principal components analysis and then pass the scored principal components to successor nodes for further analysis.

Through the interactive results browser, the user is able to view analysis results in specific formats based on the type of training analysis that is performed.

If a Dmneural network training analysis is performed, fit statistics and labels for each of the training stages can be viewed. For binary targets, a classification chart is displayed. For interval targets, a plot is created for any two variables in the data set, in addition to the predicted target variable and the residuals.

If a principal components analysis is performed, any of the following diagrams can be viewed:

- **Scaled variation.** Plots the proportional eigenvalue for each principal component.
- **Scree.** Plots the eigenvalues for successive principal components.
- **LEV.** Plots the log eigenvalues for successive principal components.
- **Cumulative.** Plots the cumulative proportional eigenvalue for each principal component.

User-Defined Model Node

The User-Defined Model node enables users to generate assessment statistics using predicted values from a model built with the SAS Code node (for example, a logistic model using the LOGISTIC procedure in SAS/STAT) or from the Variable Selection node. Additionally, the predicted values can be saved to a data set and inserted into the process flow with the Input Data Source node.

Ensemble Node

The Ensemble node enables users to combine the results from multiple models to create a single, integrated model for their data. This node performs:

- Stratified modeling.
- Bagging.
- Boosting.
- Combined modeling.

A common ensemble approach would be to re-sample the data and fit a separate model for each sample, and then combine the models to form a potentially stronger, stable model solution. Another common approach would be to combine two or more modeling methods that are fit to the same training data, such as a neural network and a decision tree. The posterior probabilities (for

class targets) or the predicted values (for interval targets) are averaged from multiple models. The ensemble model can then be used to score new data.

The models generated by the Ensemble node are fully integrated with the Model Manager and Assessment node and can be compared with other models generated by Enterprise Miner.

Memory-Based Reasoning Node

The Memory-Based Reasoning (MBR) node uses a *k*-nearest neighbor approach to categorize or predict observations. The main parameters of the MBR algorithm are the numbers of the nearest neighbors to search for and the search algorithm itself. To retrieve the nearest neighbor as fast as possible, the MBR node provides the following search algorithms, which are used to store the training data:

- Scan.
- Rd-tree (Reduced Dimensionality Tree).

The default value of *k* (the number of nearest neighbors) is 16. However, this value can be changed manually by the user.

Two Stage Model Node

The Two Stage Model node enables users to combine a classification model for a class target with a prediction model for an interval target. The interval target is usually associated with a level of the class target. For example, marketers might build a classification model to group prospects into responders and non-responders, and then build a prediction model to estimate the expected revenue for the responders.

The following models may be selected as the classification and the prediction models:

- Regression.
- Tree.
- Neural Network.

Users can also set the following Two Stage Model settings to specify how the value model is combined with the class model:

- **Transfer Function.** Specifies whether the posterior probability of the class target or the predicted classification will be used as input for generating the prediction model.
- **Filter.** Specifies the portion of the training data set to be used in training the prediction model.
- **Bias Adjustment.** Specifies how the event posterior probability from the class model and the prediction from the prediction model are combined to compute the final value prediction.

The Assessment Nodes

Assessment provides a common framework for comparing multiple models and predictions from the predictive modeling tools in Enterprise Miner. The common criteria for all modeling and predictive tools are the expected and actual profits for a project that uses the model results. These are the criteria that enable users to make cross-model comparisons and assessments, independent of all other factors, such as sample size or the type of modeling tool used.

Assessment Node

The Assessment node provides a common framework for comparing models from the Neural Network, Regression, Tree, Ensemble, User-Defined Model, Memory Based Reasoning, Princomp/Dmneural and Two Stage Model nodes.

Assessment statistics are automatically computed when a user trains a model with a modeling node. These statistics can be assessed either by using the Assessment node or the Model Manager.

An advantage of the Assessment node is that it enables users to compare models created by different modeling nodes. The Model Manager is restricted to comparing models that are trained by the respective modeling node. However, the Model Manager enables users to interactively re-define the cost or the profit values of the decision matrix for a model. This ability allows for what-if scenarios for various business situations.

The Assessment node requires a scored data set and a decision matrix. A scored data set consists of a set of predicted event probabilities. The decision matrix is a table of expected revenues and expected costs for each decision alternative for each event-level of the target variable.

The results of the Assessment node can be viewed in any of the following chart formats:

- Lift charts (or gains charts).
- Profit charts.
- Return on investment (ROI) charts.
- Cross-lift charts.
- Diagnostic classification charts.
- Statistical receiver operating characteristic (ROC) charts.
- Response threshold charts.
- Threshold-based charts.
- Interactive profit/loss assessment charts.

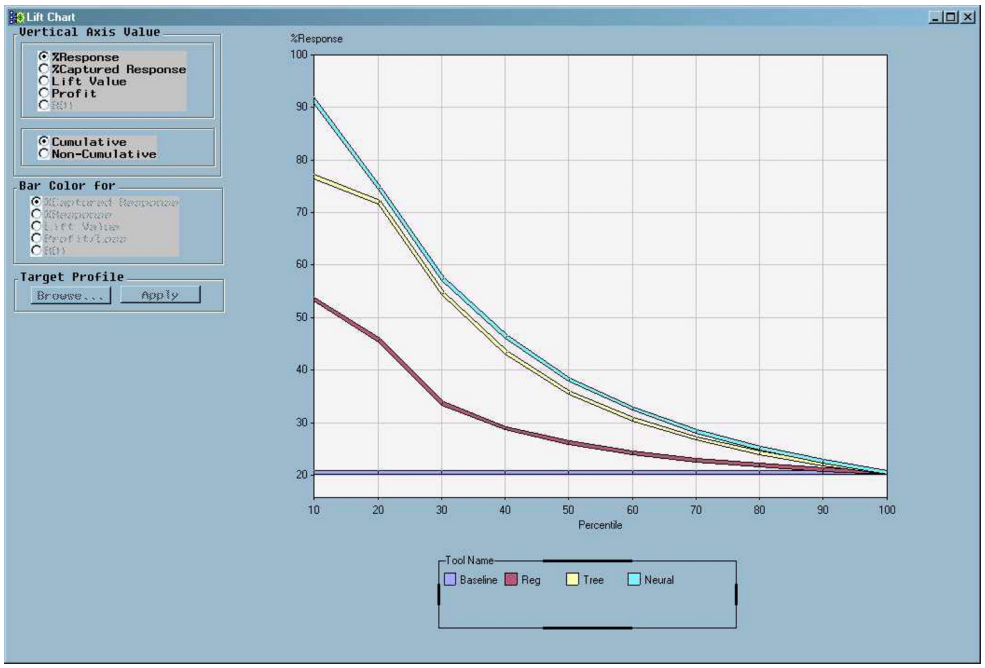

Figure 8: Lift chart that shows the results for logistic regression, decision tree and neural network models.

Reporter Node

The Reporter node enables users to assemble the results from an analysis into a single HTML report that can be viewed with a Web browser.

Each report contains descriptive information about the report, an image of the PFD, and a separate sub-report for each node in the flow. The reports are time-stamped and can be used for auditing and tracking.

The Model Deployment Nodes

Although not part of the SEMMA methodology, the model deployment nodes are still a key differentiator of Enterprise Miner. These nodes enable users to generate complete scoring code for all stages of model development, including variable reduction and transformations. Scoring code can be applied immediately to new data on any SAS system, which results in significant time savings. Scoring code is also available in the C and Java languages to enable users to deploy results outside of the SAS environment.

Score Node

The Score node enables users to manage, edit, export and execute SAS scoring code that is generated from trained models. Scoring a new data set is the end result of most data mining problems. For example:

- A marketing analyst might want to score a database to create a mailing list that contains the names of customers who are most likely to make a purchase.
- A financial analyst might want to score credit union member transactions to identify probable fraudulent transactions.

The Score node generates and manages scoring formulas in the form of a single SAS DATA step, which can be used in most SAS environments, even those that don't include Enterprise Miner. For example, you could deploy the SAS DATA step on a mainframe using only base SAS software.

C*Score Node

The C*Score node translates the SAS DATA step score code that is generated by Enterprise Miner tools into a score function in the C programming language, as described in the *ISO/IEC 9899 International Standard for Programming Languages - C handbook*. Users can save the score function to a plain text output file, which enables them to deploy the scoring algorithms in the C or C++ development environment. The C*Score node is intended for use by programmers with experience writing code in the C or C++ languages.

The C*Score node produces a header file that is used to compile the C code. The C code compiles with any C compiler that supports the *ISO/IEC 9899 International Standard for Programming Languages - C*. It can be linked as a callable C function, such as a dynamic link library (DLL).

While C*Score node output can be used as the core of a scoring system, it is not a complete scoring system. It provides only the functionality that is explicit in the SAS DATA step scoring code that is translated. Additionally, the C*Score node only translates the score code that is produced by Enterprise Miner nodes. The C*Score node cannot translate code produced by the SAS Code node.

The Utility Nodes

Although not part of the SEMMA methodology, Enterprise Miner includes a series of utility nodes that provide support features that can streamline the data mining process.

Group Processing Node

The Group Processing node enables users to define group variables, such as gender, to obtain separate analyses for each level of the grouping variable(s). If more than one target variable is defined in the Input Data Source node, then a separate analysis is performed for each target. There can be one Group Processing node per PFD. By default, group processing occurs automatically when a PFD that contains a Group Processing node is run.

The Group Processing node can also be used in combination with the Ensemble node to automatically apply bagging and boosting to any modeling algorithm.

Data Mining Database (DMDB) Node

The Data Mining Database node enables users to create a data mining database (DMDB). A DMDB is a SAS data set that is designed to optimize the performance of the analytical nodes by reducing the number of passes through the data required by the analytical engine. DMDBs contain a metadata catalog with summary statistics for numeric variables and factor-level information for categorical variables. The Data Mining Database node provides the mechanism for browsing the DMDB statistics. If a node requires a DMDB, then that node dynamically generates a DMDB when the node is run.

SAS Code Node

The SAS Code node enables users to incorporate SAS code into PFDs. As a result, the SAS Code node extends the functionality of Enterprise Miner by enabling users to integrate other SAS procedures in their data mining analysis. A SAS DATA step can also be written to create customized scoring code, to conditionally process data, and to concatenate or merge existing data sets.

The SAS Code node provides a macro facility to dynamically reference data sets and variables. Users can also compute and update a data mining database, view details about import and export data sets, and browse the code results. The SAS Code node can be defined in any stage of a PFD.

Control Point Node

The Control Point node enables users to create a control point within a PFD. Control points can reduce the number of connections required in some diagrams. For example, if multiple Input Data Source nodes connect to multiple modeling nodes, many connections are required and the PFD becomes more complex. However, the number of connections required can be reduced by connecting the Input Data Source nodes to a control point and then connecting the control point to the modeling nodes.

Subdiagram Node

The Subdiagram node can be used to represent a collection of other nodes. Subdiagrams can assist in controlling and simplifying the process flow for complex PFDs. Subdiagrams can be created by either of the following actions:

- Adding nodes directly into a Subdiagram node.
- Selecting existing nodes and connections in a diagram and condensing them into a Subdiagram node.

Client/Server Capabilities

The Enterprise Miner client/server architecture enables users to employ large UNIX servers or mainframes to access and process enormous data sources from different database management systems.

Specifically, the Enterprise Miner client/server functionality provides the following advantages:

- Distributes data-intensive processing to the most appropriate machine.
- Minimizes network traffic by processing the data on the source machine.
- Minimizes data redundancy by maintaining one central data source.
- Distributes server profiles to multiple clients.
- Enables user to regulate access to data sources.
- Enables user to toggle between remote and local processing.

Client/Server Requirements

The following table lists the client and server options available in Release 4.1 of Enterprise Miner.

Client Options	Server Options
Microsoft Windows 98	ABI+ for Intel
Microsoft Windows NT	AIX (32, 64 bit)
Microsoft Windows 2000	Compaq Tru64 UNIX (32, 64 bit)
	HP-UX (32, 64 bit)
	Intel Linux
	MVS/ESA (or prior releases) including all releases of OS/390
	Solaris (32, 64 bit)
	Microsoft Windows NT and 2000

Base SAS and SAS/STAT are required on the same server or mainframe on which the Enterprise Miner server component resides and processes data.

Conclusion

Enterprise Miner can uncover valuable information hidden in large amounts of data. This software fully integrates all steps of the data mining process beginning with the sampling of data, through sophisticated data analyses and modeling, to the dissemination of the resulting information. The functionality of Enterprise Miner is provided through a flexible GUI that enables users, who have different degrees of statistical expertise, to plan, implement, and refine their data mining projects.

References and Further Reading

Adriaans, Pieter and Dolf Zantinge. 1996. *Data Mining*. Harlow, England: Addison-Wesley.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth, Inc.

Berry, Michael J. A. and Gordon Linoff. 1997. *Data Mining Techniques*. New York: John Wiley & Sons, Inc.

Kass, G.V. 1980. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics* 29: 119-127. Standard reference for the CHAID algorithm Kass described in his 1975 Ph.D. thesis.

SAS Institute Inc. 1996. (White Paper) "Data Mining with the SAS System: from Data to Business Advantage."

SAS Institute Inc. 1997. (White Paper) "Business Intelligence Systems and Data Mining."

SAS Institute Inc. 1998. (Best Practices Paper) "Data Mining and the Case for Sampling."

SAS Institute Inc. 1999. (Best Practices Paper in Conjunction with Federal Data Corporation) "Using Data Mining Techniques for Fraud Detection."

Weiss, Sholom M. and Nitin Indurkhya. 1998. Predictive Data Mining: A Practical Guide. San Francisco: Morgan Kaufmann Publishers, Inc.

Finding the Solution to Data Mining

World Headquarters
and SAS Americas
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000
Fax: (919) 677 4444
U.S. & Canada sales:
(800) 727 0025

SAS International
PO Box 10 53 40
Neuenheimer Landstr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160
Fax: (49) 6221 474850

www.sas.com

44958US.1001