Enterprise Miner™ Software: Changes and Enhancements, Release 4.1

The Power to Know™

SAS Publishing

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *Enterprise Miner™ Software: Changes and Enhancements, Release 4.1*, Cary, NC: SAS Institute Inc., 20001.

Enterprise Miner™ Software: Changes and Enhancements, Release 4.1

Copyright © 2001 by SAS Institute Inc., Cary, NC, USA.

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, February 2001

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

IBM® and all other International Business Machines Corporation product or service names are registered trademarks or trademarks of International Business Machines Corporation in the USA and other countries.

Oracle® and all other Oracle Corporation product or service names are registered trademarks or trademarks of Oracle Corporation in the USA and other countries.

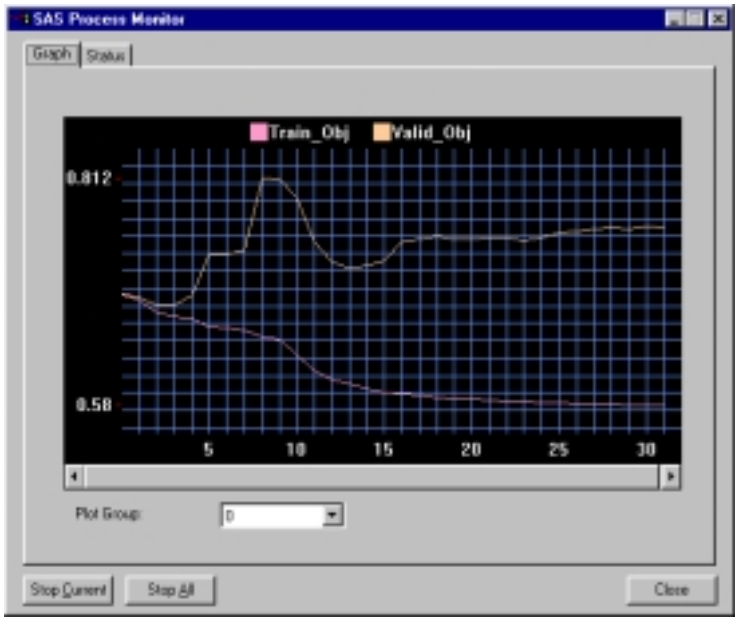
Other brand and product names are registered trademarks or trademarks of their respective companies.

Enterprise Miner™ Software: Changes and Enhancements, Release 4.1

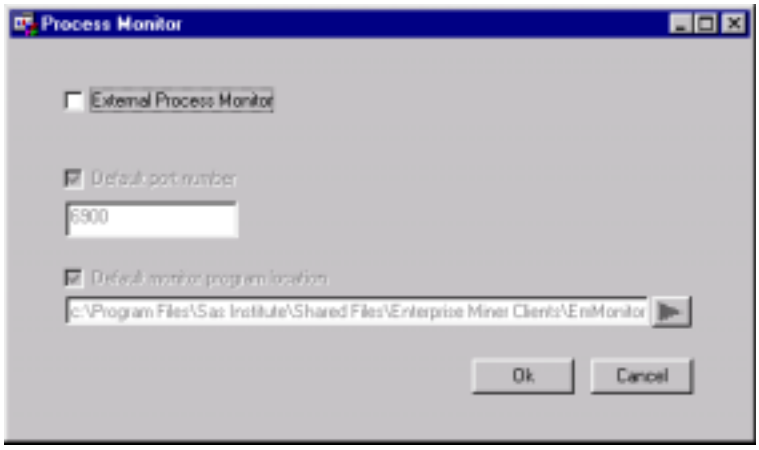
Table of Contents

SAS Process Monitor

The **Regression**, **Neural Network**, and **SOM/Kohonen** nodes use the **SAS Process Monitor** to track nodes and stop them from running. The following display shows the contents of the **SAS Process Monitor** Graph tab while training a neural network:



You can change the settings of the **SAS Process Monitor** in the Process Monitor window.



To open the Process Monitor window in the **Regression** and **SOM/Kohonen** nodes,

1. Open the node.
2. From the main menu, select **Tools→ Training Monitor Setup**.

To open the Process Monitor window in the **Neural Network** node,

1. Open the node.
2. In the General tab, select the **Configure** button next to the **Training process monitor** check box.

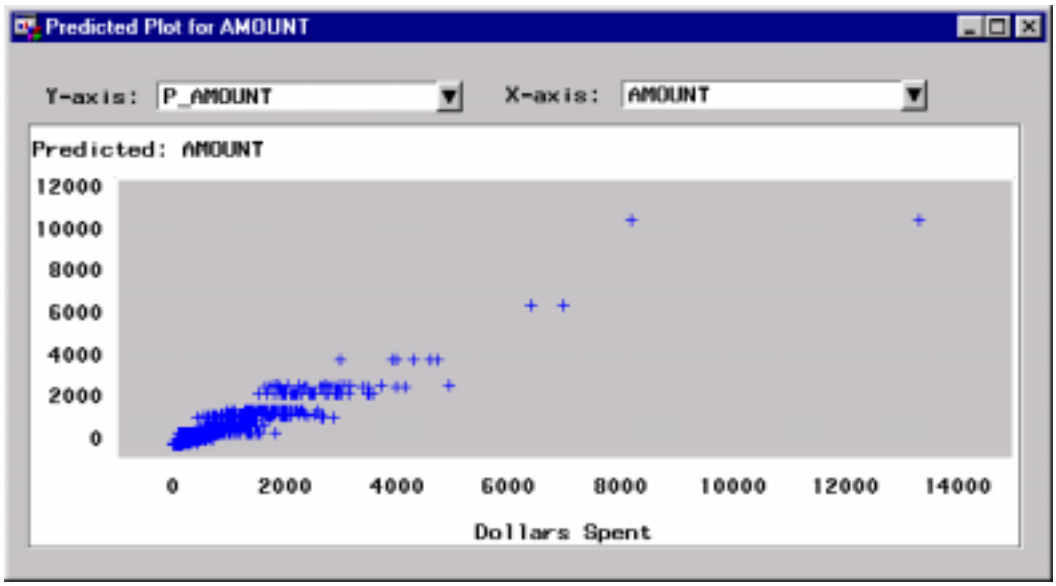
For more information about the **SAS Process Monitor**, select from the Enterprise Miner main menu **Help→ EM Reference → SAS Process Monitor**.

Summary of Node Updates

Assessment

Enhancements to the **Assessment** node and to the **Model Manager** include the following:

- A scatter plot is available, for interval targets only. The plot enables you to view the pair-wise distribution of any two variables. To create a scatter plot, select one model from the Assessment Results Browser or from the Model Manager window and then select **Tools→ Predicted Plot**. The following display shows a scatter plot of the predicted target plotted against the actual target:



- The **Export Model** is experimental. This feature enables you to export both model results and metadata and either store them in a file on the PC, or register them in a repository.

C*Score

Enhancements to the **C*Score** node include the following:

- In the C Code tab, the **Produce XML function and data description** check box enables you to create the XML-generated output at the top of the C Code tab. Also, the **Use individual input and output values as separate arguments** check box can be used to cause each data set variable in the function to be listed as a separate parameter to the function.
- The **Edit** button in the C Code tab of the C*Score Results Browser opens a PREVIEW window. The generated C code is displayed in the PREVIEW window and can be edited, printed, and saved as a separate file.

Distributions of variables can now be viewed in the Variables tab when the **Data Set Attributes** node is open.

Enhancements to the **Filter Outliers** node include the following:

- In the Settings tab, the new **Use sample** and **Use entire data** radio buttons enable you to specify the data set for the automatic and manual filtering.
- In the Settings tab, the names of the **Apply all** and **Apply unchanged** buttons have been changed to **Apply these filters to all vars** and **Apply only to vars without existing filters**, respectively, to be more descriptive.
- For class variables, you can select the values of a variable to be excluded in the Variable Histogram window by clicking a bar. The Variable Histogram window is enhanced to enable you to view the frequencies in a table format. To open a table view of frequencies of a variable, select from the main menu **View → Frequency table**. In the frequency table, select the values for exclusion.
- For interval variables, you can set the range of inclusion in the Select Value window. When the Select Value window is open, it displays the most recently saved values of **MIN** and **MAX**. If both of the displayed boundary values are within the actual data range, you cannot set the values outside the actual data range.

Enhancements to the **Regression** node include the following:

- The **Regression** node now supports nominal target variables.
- The **Minimize resource usage** check box has been added to the Advanced tab. This is useful when the model has a large number of parameters to be estimated, especially for nominal target variables.

In the General subtab of the Defaults tab, a **Replace unknown level with** check box has been added. This is useful when the class variables in the score data set contain values that are not in the training data set. You can replace the unknown value by selecting either **most frequent value (count)** or **missing value**.

Enhancements to the **Sampling** node include the following:

- Distributions of variables can be viewed in the Variables tab when the **Sampling** node is open.
- In the Stratified tab, the **Enforce proportions** check box has been added. Use it to specify what proportion of the sample will come from each stratum. Note: When a stratum of the stratified variable does not have enough observations to meet the requested proportions, selecting this check box adjusts the requested sample size to preserve the requested proportion. If this check box is not selected, the node will try to meet the sample size requirement and select as many observations as possible from each stratum. This might lead to sample proportions that are different from what you requested.

The map plot in the Map tab of the Results Browser now resembles a map, instead of a grid plot.

New Enterprise Miner Nodes and Node Features

The **Link Analysis** node is experimental. The **Link Analysis** node transforms data from differing 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 **Memory-Based Reasoning** node is experimental. The **Memory-Based Reasoning** node is a modeling tool that uses a **k**-nearest neighbor algorithm to categorize or predict observations.

Princomp/
Dmneural

The **Princomp/Dmneural** node fits an additive nonlinear model that uses the bucketed principal components as inputs to predict a binary or an interval target variable. It also performs a stand alone principal components analysis and passes the scored principal components to the successor nodes.

Time Series

The **Time Series** node is experimental. The **Time Series** node converts transactional data to time series data, and performs seasonal and trend analysis on an interval target variable.

Two Stage
Model

The **Two Stage Model** node computes a two-stage model to predict a class target and an interval target. The interval target variable is usually the value that is associated with a level of the class target variable.

Tree

The **Tree Results Viewer** is experimental. Enhancements to the **Tree Results Viewer** include the following:

- The presentation quality of the Tree view is improved.
- Abbreviated splitting rules are displayed.
- Zooming capability is improved.
- New views that display competing and surrogate splitting rules are added.
- Node colors in the Tree Results Viewer can be specified.
- Context Help has been added and can be accessed by pressing the F1 key.

Experimental Text-Mining Nodes

The following experimental text-mining nodes require a separate setinit. For more information, contact your SAS sales representative.

Text Parsing

The **Text Parsing** node decomposes textual data into term and document frequency tables that are suitable for data mining. The input data is a collection of documents or text that you want to examine for patterns or that you want to mine for information.

SVD

The **Singular Value Decomposition (SVD)** node provides a technique to reduce the dimensions from thousands to one or two hundred. In text-mining applications, it resolves the problem of determining when two documents are similar by performing a latent semantic analysis, in which terms and documents that are closely associated are placed near one another. Also, documents that have similar conceptual meaning but dissimilar term patterns are generally placed close together.

E-M
Clustering

The **Expectation-Maximization Clustering (E-M Clustering)** node performs observation clustering by identifying primary and secondary clusters for a given data set. In the **Expectation-Maximization Clustering** node, observations are not grouped into clusters, but belong to each primary cluster with a certain probability.