Products and Solutions

- Industries
- Solution Lines
- Data Integration
- Business Intelligence
- Analytics
  - Data & Text Mining
    - Predictive Analytics/Data Mining
    - Text Mining
  - Forecasting & Econometrics
  - Operations Research
  - Quality Improvement
  - Statistics
- Enterprise Intelligence Platform
- Government & Education
- Product Index A-Z

Enterprise Miner™
SEMMA

The acronym SEMMA -- sample, explore, modify, model, assess -- refers to the core process of conducting data mining. Beginning with a statistically representative sample of your data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy.

Before examining each stage of SEMMA, a common misunderstanding is to refer to SEMMA as a data mining methodology. SEMMA is not a data mining methodology but rather a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining. Enterprise Miner can be used as part of any iterative data mining methodology adopted by the client. Naturally steps such as formulating a well defined business or research problem and assembling quality representative data sources are critical to the overall success of any data mining project. SEMMA is focused on the model development aspects of data mining:

- Sample (optional) your data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. For optimal cost and performance, SAS Institute advocates a sampling strategy, which applies a reliable, statistically representative sample of large full detail data sources. Mining a representative sample instead of the whole volume reduces the processing time required to get crucial business information. If general patterns appear in the data as a whole, these will be traceable in a representative sample. If a niche is so tiny that it's not represented in a sample and yet so important that it influences the big picture, it can be discovered using summary methods. We also advocate creating partitioned data sets with the Data Partition node:
  - Training -- used for model fitting.
  - Validation -- used for assessment and to prevent over fitting.
  - Test -- used to obtain an honest assessment of how well a model generalizes.

- Explore your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration doesn't reveal clear trends, you can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. For example, in data mining for a direct mail campaign, clustering might reveal groups of customers with distinct ordering patterns. Knowing these patterns creates opportunities for personalized mailings or promotions.

- Modify your data by creating, selecting, and transforming the variables to focus the model selection process. Based on your discoveries in the exploration phase, you may need to manipulate your data to include information such as the grouping of customers and significant subgroups, or to introduce new variables. You may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. You may also need to modify data when the "mined" data change. Because data mining is a dynamic, iterative process, you can update data mining methods or models when new information is available.

- Model your data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome. Modeling techniques in data mining include neural networks, tree-based models, logistic models, and other statistical models -- such as time series analysis, memory-based reasoning, and principal components. Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data. For example, neural networks are very good at fitting highly complex nonlinear relationships.

- Assess your data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, you can test the model against known data. For example, if you know which customers in a file had high retention rates and your model predicts retention, you can check to see whether the model selects these customers accurately. In addition, practical applications of the model, such as partial mailings in a direct mail campaign, help prove its validity.

By assessing the results gained from each stage of the SEMMA process, you can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.

Once you have developed the champion model using the SEMMA based mining approach, it then needs to be deployed to score new customer cases. Model deployment is the end result of data mining - the final phase in which the ROI from the mining process is realized. Enterprise Miner automates the deployment phase by supplying scoring code in SAS, C, Java, and PMML. It not only captures the code for of analytic models but also captures the code for preprocessing activities. You can seamlessly score your production data on a different machine, and deploy the scoring code in batch or real-time on the Web or in directly in relational databases. This results in faster implementation and frees you to spend more time evaluating existing models and developing new ones.

More on this topic

- News and Events
- Fact Sheet