References

[1] Edwards,P.I., Murray, A.F., Papadapoulos, G., Gordon, M.F., Wallace, A.R. and Barnard, J. Paper curl prediction – neural networks applied to the papermaking industry. **Artificial Neural Networks Conference** (1999): 335–340.

[2] Bortolin, G., Gutman, P.O., Nilsson, B. On modeling of curl in multi-ply paperboard. **Journal of Process Control** 16(2006): 419–429.

[3] Achiche, S., Baron, L., Balazinski, M. Predictive Fuzzy Control of Paper Quality. **Fuzzy Information Processing Society** (2006): 31-34.

[4] Elo, P., Saarinen, J., Kaski, K., Pakarinen, P., Kiiskinen, H., Kaijaluoto, S., Edelmann, K. Analysis of quality properties in paper drying with neural networks. **Proceedings of International Joint Conference on Neural Networks** (1993): 1845-1848.

[5] Bissessur, Y., Martin, E.B., Morris, A.J. Monitoring the performance of the paper making process. **Control Engineering Practice** (1999): 1357-1368.

[6] Michael, J.I. **Neural Networks**, A. Tucker, (Ed.), CRC handbook of Computer Science, CRC Press, Boca Raton, FL (1996).

[7] Simon, H. **Neural Networks, a comprehensive foundation**. Pearson Education, Inc (2001).

[8] Martin, H.T. **Neural Network Design**. PWS Publishing Company (1996).

[9] Warren, S. **Neural Network FAQ**. URL: ftp://ftp.sas.com/pub/neural/FAQ.html (2003).

[10] Warren, S. Neural Networks and Statistical Models. **Proceedings of the Nineteenth Annual SAS Users Group International Conference** (1994).

[11] Matignon, R. **Neural Network Modeling**. EAuthorhouse (2005).

[12] Jonathon, S. **A Tutorial on Principal Component Analysis** 2(2005).

[13] Randall, M. **Data Mining Using SAS Enterprise Miner**. John Wiley & Sons, Inc (2007).

[14] Han, J.W., Kamber, M. **Data Mining Concepts and Techniques**. 2nd Ed., Elsevier Inc. 2006: (71-74, 363-366).

[15] Warren, S. Stopped Training and Other Remedies for Over-fitting. **Proceedings of**

the 27th Symposium on the Interface (1995).

[16] Witten, I.H., Frank, E. **Data Mining-Practical Machine Learning Tools and Techniques**. 2nd Ed., Elsevier Inc. (2005):152-153.

[17] Mithat G. Receiver Operating Characteristic (ROC) Curves. **SUGI 31 Proceedings** (2006).

APPENDICES

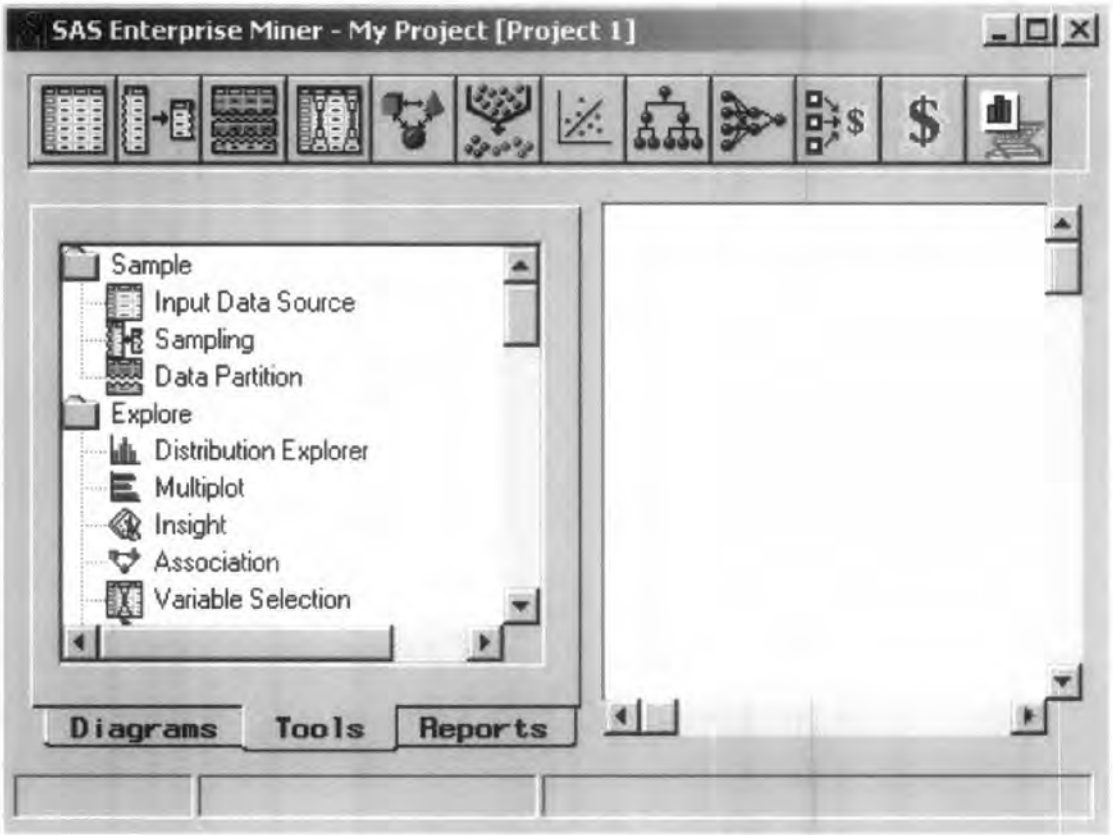
Appendix A

SAS Enterprise Miner and SEMMA Architecture

SAS Enterprise Miner is part of the SAS Institute, Inc. product line. The SAS System is an integrated system of software providing complete control over data access, management, analysis and presentation (SAS Institute, Inc, 1990). SAS is also a programming language that can be used as a tool to perform statistical functions (Speed, 2005). Many different products are included in the SAS System. The one primarily used in this project was SAS Enterprise Miner.

SAS Enterprise Miner is a tool within SAS created for data mining. Data Mining is the process of extricating knowledge from very large data sets (Tretter, 2003). According to SAS Institute, Inc., "SAS Enterprise Miner streamlines the entire data mining process from data access to model deployment by supporting all necessary tasks within a single, integrated solution, all while providing the flexibility for efficient workgroup collaborations." (SAS Institute, Inc, 2006) In general, SAS Enterprise Miner can be used to take large amounts of information and display the information in a way that it can be used and interpreted, the interface of EM is shown in the following figure A.1.

SAS Institute defines data mining as the process of Selecting, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage. The data mining process is applicable across a variety of industries and provides methodologies for such diverse business problems as fraud detection, house holding, customer retention and attrition, database marketing, market segmentation, risk analysis, affinity analysis, customer satisfaction, bankruptcy prediction, and portfolio analysis.



A.1: Enterprise Miner 4.3 Interface.

Enterprise Miner software is an integrated product that provides an end-to-end business solution for data mining. A graphical user interface (GUI) provides a user-friendly front-end to the SEMMA data mining process:

- **Sample** the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- **Modify** the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.

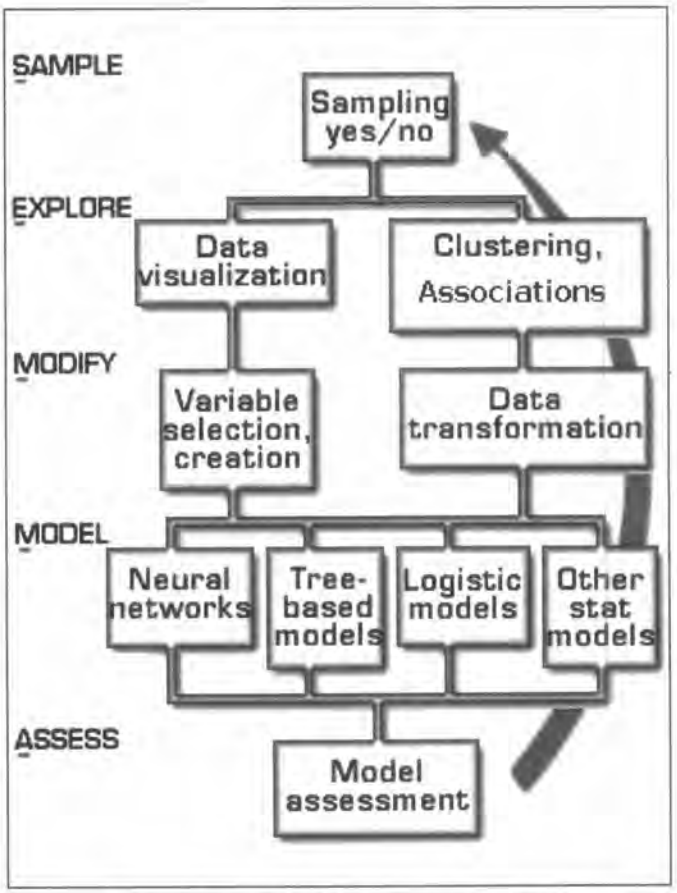


Figure A.2: SEMMA process illustration.

You may or may not include all of these steps in your analysis, and it may be necessary to repeat one or more of the steps several times before you are satisfied with the results. After you have completed the assess phase of the SEMMA process detailed in Figure A.2, you apply the scoring formula from one or more champion models to new data that may or may not contain the target. Scoring new data that is not available at the time of model training is the end result of most data mining problems.

The SEMMA data mining process is driven by a process flow diagram, which you can modify and save. The GUI is designed in such a way that the business analyst who has little statistical expertise can navigate through the data mining methodology, while the quantitative expert can go "behind the scenes" to fine-tune and tweak the analytical process.

Enterprise Miner contains a collection of sophisticated analysis tools that have a common user-friendly interface that you can use to create and compare multiple models. Statistical tools include clustering, self-organizing maps/Kohonen, variable

selection, trees, linear and logistic regression, and neural networking. Data preparation tools include outlier detection, variable transformation, data imputation, random sampling, and the partitioning of data sets (into training, testing, and validation data sets). Advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare modeling results.

Appendix B

Coding

SAS Language has been used in program coding. Related codes in this thesis have been shown as follows:

```
/*** Data Partition (Random Sample) ***/

  %let seed = 12345;
  data
    EMDATA.TRNFIAQO
    EMDATA.VALID814
    EMDATA.TESTSSD
  ;
  drop _c00:
  ;
  set EMDATA.VIEW_OUE;
    if (2280 +1-_n_)*ranuni(12345) <= (1368 - _c000001) then do;
      _c000001 + 1;
      output EMDATA.TRNFIAQO;
    end;
    else  if (1367+1-_n_)*ranuni(12345) <= (228 -_c000002) then do;
      _c000002 + 1;
      output EMDATA.VALID814;
    end;
    else do;
      _c000003 +1;
      output EMDATA.TESTSSD;
      end;
    run;
```

```
/*** standardizing the input variables from a standard normal dist with mean of one
and a variance of one ***/

proc standard data=dmddata out=standardout mean=0 std=1;
  var x;
run;

/*** calculating the ROC Curve ***/
proc rank data=b.roc out=b.roc1;
  var predict;
  ranks rpredict;
run;

proc sql;
select sum(target=1) as n1,
      (sum(rpredict*(target=1))-.5*(calculated n1)*(calculated n1+1))
      /((calculated n1)*(count(target)-(calculated n1))) as c
      from b.roc1;
quit;


/*** Evaluation for sensitivity and specificity ***/

data a;
  input predict $ target $ weight;
  datalines;
 accept accept 616
 reject accept 4
 accept reject 21
 reject reject 43
  run;
  proc freq data = a;
```

```
table predict*target;
weight weight;
run;
proc freq data = a;
by target;
table predict / binomial;
weight weight;
run;

/*** Dmneural training code for principal components analysis ***/
proc dmneurl data=EMDATA.VIEW_744 dmdbcat=EMPROJ.dm_DGM00025
outstat=EMPROJ.STA_UR22 outclass=EMPROJ.CLA_X7ZX
CORR;
var /*------ input variables ---------*/
...
...
...
;
run;

/*** Variable Selection R-Square ***/
proc dmine data=_EMSPDE.sp_DGM00113 dmdbcat=EMPROJ.sp_DGM00113
minr2=0.005 stopr2=0.0005 NOMONITOR NOAOV16 NOINTER PSHORT USEGROUPS
outest=EMPROJ.OUTE2ZPK;
var
...
...
...
;
ordinal
```

```
...
...
;
target DEBTINC;
run;


/*** Data mining SAS procedure is automatically executed before the neural network
procedure to create a data mining SAS DMDB data set and catalog ***/
proc dmdb data=sasdata out=dmddata dmdbcat=dmcdata;
var x y;
run;


/*** Neural network SAS procedure with interval target var with MLP algorithm; ***/
proc neural data=dmddata dmdbcat=dmcdata graph ranscale=.1 random=0;
                /* read data mining data set */
input  x  / level=int;    /*  standardized input interval variable*/
target y  / level=int error=normal std=std;
                /* normal distribution error function */
                /* standardized target interval variable */
archi mlp hidden=10      /*  MLP design with ten hidden units */
hidden =1;                /*  one hidden layer */
initial inest=inest;      /*  data set with starting values of initial weight estimates */
prelim 5 maxiter=20;      /*  5  preliminary runs w/ a max of 20 iterations */
train tech = quanew ;     /* Quasi-Newton optimization method */
outest= outestds estiter=1 /*  Parameter estimates with training and
                               validation average error statistic at each
                               iteration since estiter > 0 */
outfit = outfit;          / * model assessment statistics at ea. iteration;*/
```

```
code 'c:\temp\code.sas';
score data=sasdata out=pred nodmdb;
run;
```

VITAE

Feifei Wang was born in December 17th, 1982, in Jiangsu, China. She obtained her Bachelor's Degree in Computer Science from the Faculty of Information Science and Technology Management, Nanjing University of Technology in 2006.