

Comparison of Enterprise Miner and SAS/Stat for Data Mining

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

There are many definitions of data mining: the discovery of spurious relationships in the data; automated data analysis; examination of large data sets; exploratory data analysis. Early methods of data mining included stepwise regression, cluster analysis, and discriminant analysis. Data mining should be thought of as a process that includes data cleaning, investigation of the data using models, and validation of potential results. In particular, practical significance should be emphasized over statistical significance. Also, decision making and the cost of misclassification is important. SAS/Stat contains methods that can be used to investigate data using a data mining process. These methods can complement those developed specifically for Enterprise Miner, and can be used in conjunction with Enterprise Miner. This paper will examine data mining in SAS/Stat, contrasting it with Enterprise Miner.

INTRODUCTION

The term "data mining" has now come into public use with little understanding of what it is or does. In the past, statisticians have thought little of data mining because data were examined without the final step of model validation. Data mining differs from standard statistical practice in that the process

- Assumes large data samples
- Assumes large numbers of variables
- Has Validation as a routine, automatic part of the process
- Examines the cost of misclassification
- Emphasizes decision-making over inference

And yet there remains overlap between statistics and data mining in both technique and practice. Since many of the techniques, such as cluster analysis, overlap both statistics and data mining, it is the purpose of this paper to examine some of the differences and similarities in the use of these overlapping techniques from a statistical perspective versus a data mining perspective.

To demonstrate the different techniques, a dataset consisting of responses to a survey concerning student expectations in mathematics courses was used throughout. A list of questions asked in the survey is given in the appendix. The data were coded ordinally and variable names were coded to represent the questions. The data analysis was performed to investigate the relationship between variables and responses in the dataset. A total of 192 responses were collected from this survey. The basic survey questions are

The areas of mathematics which interest me are (Check all that apply):

- Pure mathematics
- Applied mathematics
- Abstract Algebra
- Number Theory
- Topology
- Real analysis
- Discrete mathematics
- Differential equations
- Actuarial Science
- Probability
- Statistics

The students were also asked whether they were graduate or undergraduate, and how many hours per week they estimated they studied mathematics.

CLUSTER ANALYSIS

As a data mining process, clustering is considered to be unsupervised learning, meaning that there is no specific outcome variable. Clustering involves the following process:

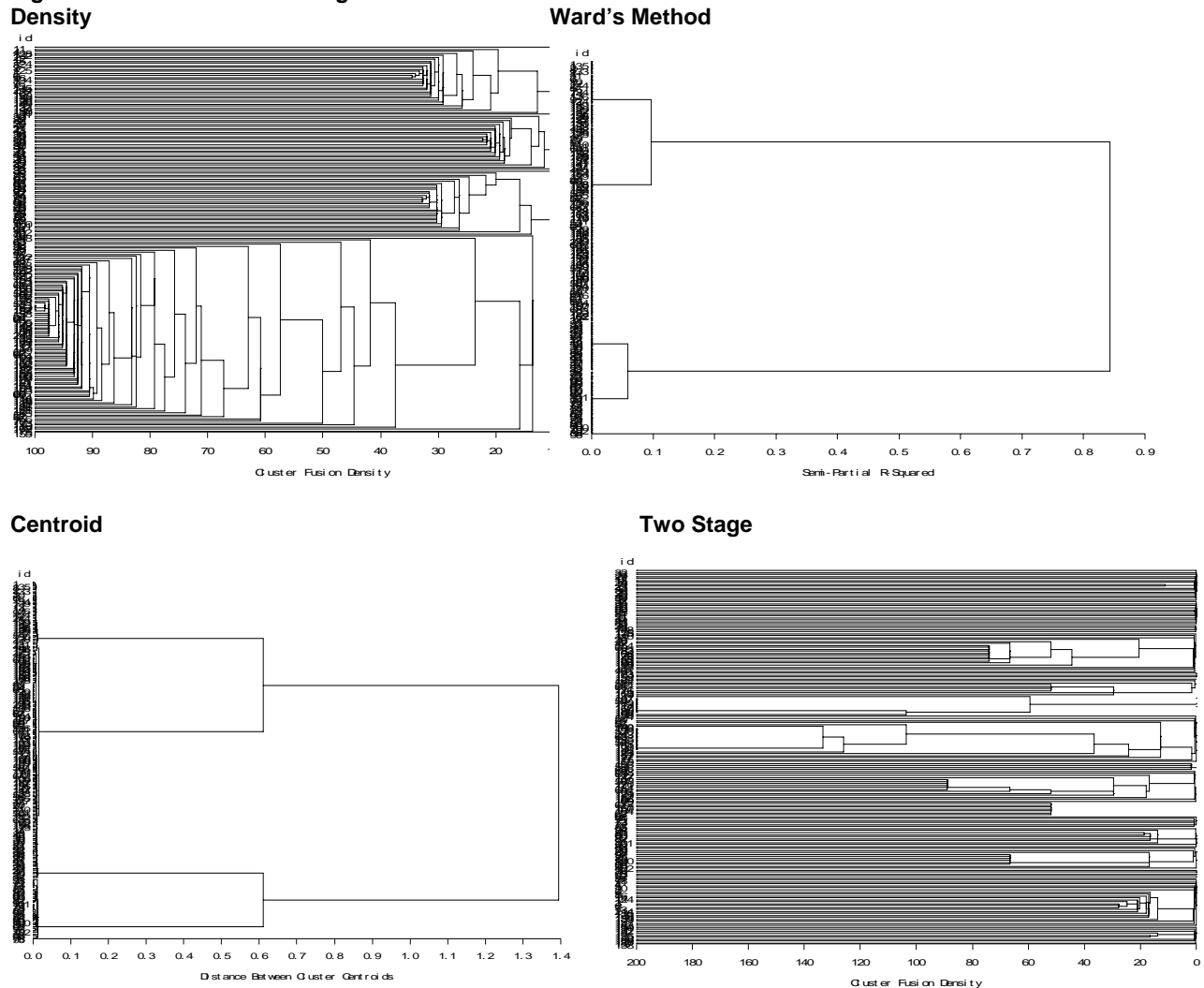
- Group observations or group variables
- Choice of type-hierarchical or non-hierarchical
- Number of clusters
- Cluster identity
- Validation

Hierarchical clustering involves the grouping of observations based upon the distance between observations. Distance can be defined using different criteria available in PROC CLUSTER. The clusters are built based on a hierarchical tree structure. The closest observations are grouped together initially followed by the next closest match. The other method of clustering is based upon the selection of random seed values as the centers of spheres containing all observations closest to that center (PROC FASTCLUS). Then the centers are re-defined based upon the outcomes. In Enterprise Miner, PROC FASTCLUS is used to perform clustering. This is the same procedure available in SAS/Stat. SAS/Stat has the additional hierarchical clustering techniques available. The variables in the dataset dealing with preferences for mathematics subject were first clustered in SAS/Stat using the hierarchical procedure in PROC CLUSTER. Several distance criteria were used for comparison purposes. In addition, course level (200,300,400, 500 and above) was included.

```
proc cluster data=sasuser.studentsurvey method=density r=2 ;
var courselevel pure applied abstractalgebra numbertheory topology realanalysis
discretemathematics differentialequations actuarialscience probability statistics;
id id;
run;
proc tree horizontal spaces=2;
id id;
```

In addition to density, Ward's method, centroid, and two stage methods were used to determine clusters with very different results (Figure 1).

Figure 1. Methods of Clustering



The number of clusters defined range from 2 to many. With so many different possible clustering results, it is difficult to make a judgment as to which is optimal. Instead, the question must become, "Is the result reasonable, and does the result enhance decision-making?"

In contrast, k-means clustering, which is used in Enterprise Miner, uses the following code:

```
proc fastclus data=sasuser.studentsurveyimputed maxclusters=4 list out=sasuser.fastclusresults ;
var courselevel pure applied abstractalgebra numbertheory topology realanalysis
discretemathematics differentialequations actuarialscience probability statistics;
id id;
```

It is important to save the cluster values. The option out=sasuser.fastclusresults preserves the original data plus the variables CLUSTER and DISTANCE. CLUSTER gives the cluster number to the observation; DISTANCE gives the Euclidean distance from the observation to the center of the cluster.

Note that the investigator needs to specify the maximum number of clusters to be examined. It is possible to create a list identifying which observation is classified into which cluster. A labeled list is not possible with the hierarchical procedure since the final clustering is uncertain. Once the hierarchy is established and the number of clusters is chosen, the final clusters can be labeled by the investigator. For the Fastclus results, Table 1 gives the means of each variable by cluster.

Table 1. Cluster by Variable

Cluster	Pure	Applied	Algebra	Topology	Probability	Statistics
1	.68	.36	.32	.02	.18	.06
2	.28	.92	.05	.08	.43	.69
3	.36	.08	.08	.04	.88	.92
4	1.00	1.00	.76	.56	.41	.35

In this program, the maximum number of possible clusters is given by the user. To validate the results, it is extremely important to give each cluster a meaningful identity. Given the proportion of values in each cluster, it is possible to define labels for each cluster (Table 2).

Table 2. Cluster Labels

Cluster	Label
1	Pure Math
2	Applied Math
3	Statistics
4	Math Generally

Five clusters are summarized in Table 3.

Table 3. 5 Clusters in FASTCLUS

Cluster	Pure	Applied	Algebra	Topology	Probability	Statistics
1	.36	.46	.07	.03	.09	.04
2	.30	.80	.48	.14	.22	.32
3	.05	.44	.02	.05	.65	.98
4	1.00	.50	.40	.05	1.00	.75
5	1.00	.90	.70	.60	.60	.40

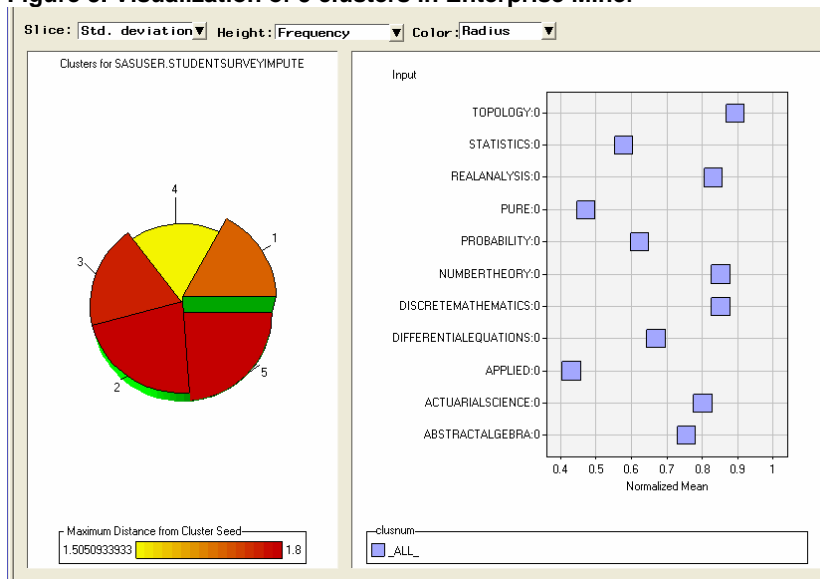
Cluster	Label
1	Pure Math
2	Applied Math
3	Statistics
4	Somewhat Ambivalent
5	Math Generally

In contrast, Enterprise Miner does not require a specification of a maximum number of clusters, although it is an available option. It also provides some graphics to examine the data that are not readily available in SAS/Stat (Figure 3). The coding for Enterprise Miner is given in Figure 2. Although defined using icons, the Fastclus procedure is the same.

Figure 2. Enterprise Miner Clustering

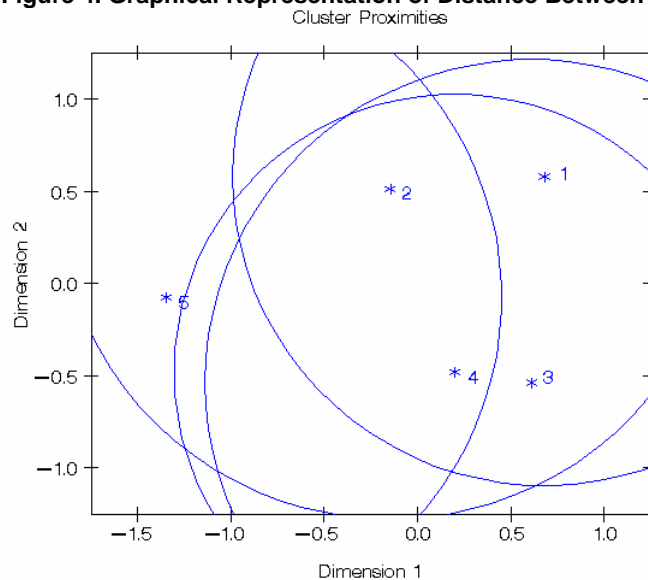


Figure 3. Visualization of 5 clusters in Enterprise Miner



In addition, a multidimensional scaling (PROC MDS) is performed in Enterprise Miner to examine the separation between clusters (Figure 4). The scaling is based upon distance within and between clusters.

Figure 4. Graphical Representation of Distance Between Clusters



The centers of each cluster are well separated but there is considerable overlap. Once the clusters are defined, specific variables can be examined (Table 4).

Note also in Figure 2 that Enterprise Miner will automatically divide the dataset, creating a holdout sample that can be used to test the results. That capability does not automatically exist in SAS/Stat. It is useful to have such a holdout sample so that the clustering results can be validated. To create such a sample, the following code is used:

```

PROC SQL;
    CREATE VIEW WORK.SORT6649
    AS SELECT * FROM sasuser.REVISEDSTUDENTSURVEY;
QUIT;

PROC SURVEYSELECT DATA=sasuser.surveysample
    OUT=SASUSER.RAND8956(LABEL="Random sample of sasuser.revisedstudentsurvey")
    METHOD=SRS
    RATE=0.2
    ;

RUN;

data sasuser.surveysampleadd;
set sasuser.surveysample;
sample=1;
run;
    
```

Creates the random sample of 20% of the data.

Adds a pointer to the random sample, of numeric value 1.

```

proc sort data=Sasuser.Revisedstudentsurvey out=WORK._TABLE1_;
  by ID;
run;
proc sort data=Sasuser.Surveysampleadd out=WORK._TABLE2_;
  by ID;
run;
data sasuser.merged;
  length id 8 Student_Type $ 16 CourseLevel 8 Course 8 Section 8 Pure 8
    Applied 8 AbstractAlgebra 8 NumberTheory 8 Topology 8 RealAnalysis 8
    DiscreteMathematics 8 DifferentialEquations 8 ActuarialScience 8
    Probability 8 Statistics 8 sample 8;
  merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
  by ID;
run;

```

Merges the random sample into the original dataset.

PROC SQL;

```

CREATE TABLE SASUSER.studentsurveytrain AS SELECT  studentsurveymerged.id
FORMAT=BEST12.,
studentsurveymerged.Student_Type FORMAT=$F16.,
studentsurveymerged.Pure FORMAT=BEST12.,
studentsurveymerged.Applied FORMAT=BEST12.,
studentsurveymerged.AbstractAlgebra FORMAT=BEST12.,
studentsurveymerged.NumberTheory FORMAT=BEST12.,
studentsurveymerged.Topology FORMAT=BEST12.,
studentsurveymerged.RealAnalysis FORMAT=BEST12.,
studentsurveymerged.DiscreteMathematics FORMAT=BEST12.,
studentsurveymerged.DifferentialEquations FORMAT=BEST12.,
studentsurveymerged.ActuarialScience FORMAT=BEST12.,
studentsurveymerged.Probability FORMAT=BEST12.,
studentsurveymerged.Statistics FORMAT=BEST12.,

```

Filters out the random sample leaving all values not in the sample.

```

studentsurveymerged.sample FORMAT=BEST12.
FROM sasuser.studentsurveymerged AS studentsurveymerged
WHERE studentsurveymerged.sample NOT = 1;
QUIT;

```

In this code, SURVEYSAMPLEADD is used to test the data; STUDENTSURVEYTRAIN is used to define the model. Once the clusters are defined using the training set, PROC DISCRIM can be used to examine the clusters that would be defined on the holdout sample:

```

proc discrim data=sasuser.studentsurveytrain testdata=sasuser.surveysampleadd list testlist;
class student_type;
id id;
var Pure Applied AbstractAlgebra NumberTheory RealAnalysis DiscreteMathematics DifferentialEquations
ActuarialScience Probability Statistics;;
run;

```

In the above code, data=sasuser.studentsurvey train is used to define the discriminant function (which will fit the training set almost perfectly). Test data (defined by testdata=sasuser.surveysampleadd) are used to define cluster values on a set not used to define them. Once these are defined, cluster means can be computed on the test data set and compared to those in the initial set. Once the clusters are defined, they can be used in other types of analyses for comparison purposes. Table 4 gives a chi-square analysis to examine clusters by types of students.

Table 4. Clusters by Student Type

Cluster	Undergraduate	Graduate
1	53 (41%)	14 (23%)
2	35 (27%)	15 (25%)
3	20 (15%)	23 (38%)
4	16 (12%)	4 (7%)
5	5 (5%)	4 (7%)

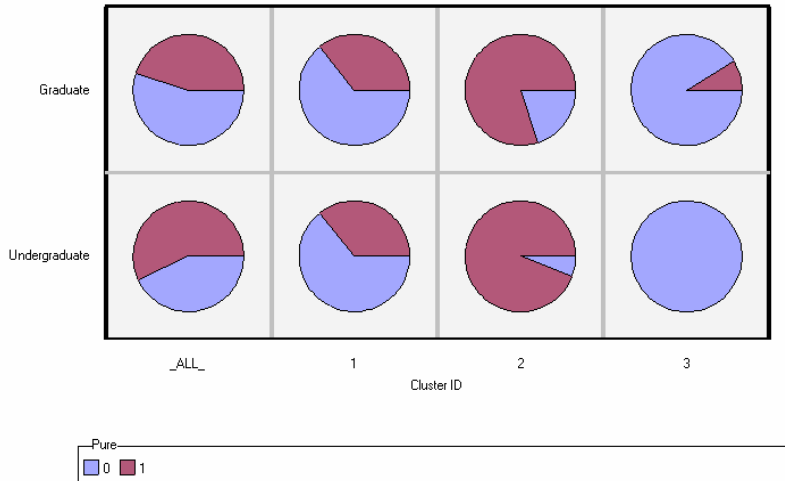
Undergraduates are more likely than graduate student to prefer pure mathematics. This trend is even more noticeable when clusters are examined by course level (Table 5).

Table 5. Clusters by Course Level

Cluster	200	300	400	500
1	41 (42%)	12 (36%)	6 (21%)	8 (25%)
2	24 (25%)	11 (33%)	11 (39%)	4 (13%)
3	14 (14%)	6 (18%)	8 (28%)	15 (47%)
4	13 (13%)	3 (9%)	2 (7%)	2 (6%)
5	5 (5%)	1 (3%)	1 (4%)	3 (9%)

In contrast to tables, Enterprise Miner allows this information to be depicted graphically in the CLUSTER NODE (Figure 5).

Figure 5. Enterprise Miner Comparisons



CLASSIFICATION AND PREDICTIVE MODELING

Classification is considered to be supervised data mining since it is possible to compare a predicted value to an actual value. In SAS/Stat, discriminant analysis and logistic are the primary means of classification. In Enterprise Miner, neural networks, decision trees, and logistic regression are used. However, the two components of SAS approach classification with different perspectives. In Enterprise Miner, the dataset is large enough to partition it so that the classification model can be validated through the use of a holdout sample. Misclassification is one of the means of determining the strength of the model. Another is to define a profit/loss function to determine the cost (or benefit) of a correct or incorrect classification.

In SAS/Stat, datasets are often small so that partitioning is not possible. The strength of a logistic regression is measured by the odds ratios, the receiver operating curve, and the p-value. Similarly, the strength of discriminant analysis is measured by the proportion of correct classifications without the use of a holdout sample, although there is an option for cross validation that is not available for logistic regression.

For logistic regression, however, the two data sets should remain merged. However, the coding given in the previous section should be modified to allow for differences in procedures.

```
data sasuser.surveysampleadd;
set sasuser.surveysample;
sample=1;
student_type1=student_type;
student_type='.';
run;
```

```
proc sort data=Sasuser.Revisedstudentsurvey out=WORK._TABLE1_
by ID;
run;
proc sort data=Sasuser.Surveysampleadd out=WORK._TABLE2_
by ID;
run;
data sasuser.merged;
length id 8 Student_Type $ 16 student_type1 $ 16 CourseLevel 8 Course 8 Section 8 Pure 8
Applied 8 AbstractAlgebra 8 NumberTheory 8 Topology 8 RealAnalysis 8
```

Merges the random sample into the original dataset.

```

DiscreteMathematics 8 DifferentialEquations 8 ActuarialScience 8
Probability 8 Statistics 8 sample 8;
merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
by ID;
run;

```

Then, using the merged datasets, the outcome values are removed from the smaller partition. The logistic procedure will predict the values that can then be compared to accuracy. In this example, the holdout sample had a 22% misclassification rate compared to 13% for the initial sample.

Similarly, PROC DISCRIM uses the following code, where the partitioned dataset can be used as a holdout sample. Note that it is very similar to the code used to validate the clustering in the previous section.

```

proc discrim data=sasuser.trainingset testdata=sasuser.partition list testlist;
class student_type;
id id;
var pure applied abstractalgebra numbertheory topology realanalysis discretemathematics
differenialequations actuarialscience probability statistics;
run;

```

The summary table on the test data is given in Table 6 below.

Student Type	Graduate	Undergraduate
Graduate	4 (67%)	2 (33%)
Undergraduate	6 (30%)	14 (70%)

Nonparametric discriminant analysis can also be used:

```

proc discrim data=sasuser.trainingset testdata=sasuser.partition list testlist
method=npair r=1 kernel=epa;
class student_type;
id id;
var pure applied abstractalgebra numbertheory topology realanalysis discretemathematics
differenialequations actuarialscience probability statistics;
run;

```

with a strong likelihood of correctly classifying undergraduates (Table 7).

Student Type	Graduate	Undergraduate
Graduate	3 (50%)	3 (50%)
Undergraduate	4 (20%)	16 (80%)

Instead of a holdout sample, cross validation is the standard method to correct for the inflation of results (Table 8):

```

proc discrim data=sasuser.revisedstudentsurvey list
method=npair k=5 crossvalidate;
class student_type;
id id;
var pure applied abstractalgebra numbertheory topology realanalysis discretemathematics
differenialequations actuarialscience probability statistics;
run;

```

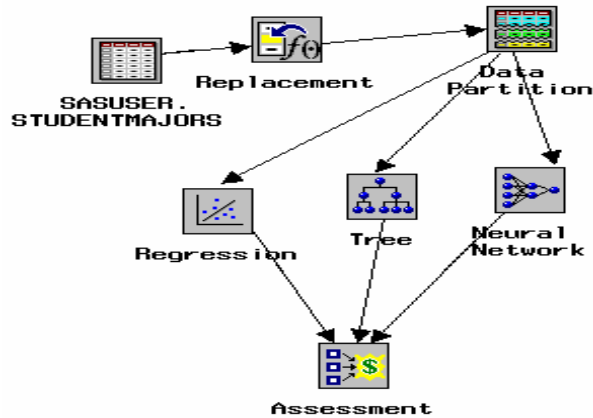
Student Type	Graduate	Undergraduate
Graduate	28 (47%)	31 (52%)
Undergraduate	46 (35%)	83 (64%)

Note that while the cross validation technique can correctly classify undergraduates 64% of the time, the classification of graduate students is poor. In comparison, a standard discriminant analysis without cross validation shows a much more accurate result (Table 9).

Student Type	Graduate	Undergraduate
Graduate	43 (72%)	17 (28%)
Undergraduate	43 (33%)	87 (67%)

In contrast, Enterprise Miner has the capability to partition easily (Figure 6).

Figure 6. Code in Enterprise Miner for Classification



Neural networks act like “black boxes” in that the model is not presented in a nice, concise format that is provided by regression. Its accuracy is examined in a way similar to the diagnostics of the regression curve. The simplest neural network contains a single input (independent variable) and a single target (dependent variable) with a single output unit. It increases in complexity with the addition of hidden layers, and additional input variables. With no hidden layers, the results of a neural network analysis will resemble those of regression. Each input variable is connected to each variable in the hidden layer, and each hidden variable is connected to each outcome variable. The hidden units combine inputs, and apply a function to predict outputs. Hidden layers are often nonlinear.

Decision trees provide a completely different approach to the problem of classification. The decision tree develops a series of if...then rules. Each rule assigns an observation to one segment of the tree, at which point there is another if...then rule applied. The initial segment, containing the entire dataset, is the root node for the decision tree. The final nodes are called leaves. Intermediate nodes (a node plus all its successors) forms a branch of the tree. The final leaf containing an observation is its predictive value. Unlike neural networks and regression, decision trees will not work with interval data. It will work with nominal outcomes that have more than two possible results. Decision trees will also work with ordinal outcome variables.

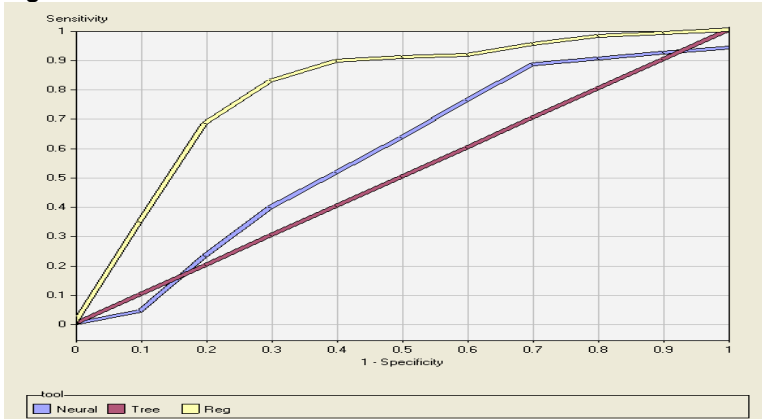
The Assessment Node is used to compare outcome methods (Table10).

Table 10. Comparison of Methods

Tool	Target Event	Misclass Rate	Valid: Misclass Rate	Test: Misclass Rate
Neural Network	Undergraduate	0.381578947 4	0.333333333 3	0.245614035 1
Tree	Undergraduate	0.342105263 2	0.315789473 7	0.280701754 4
Regression	Undergraduate	0.197368421 1	0.473684210 5	0.438596491 2

Interestingly, logistic regression has the lowest misclassification rate on the initial training set but the highest misclassification rate on the testing set. The results suggest that logistic regression tends to inflate results. Similarly, the receiver operating curve is given in Figure 7 for three models. Note also that the data sample is partitioned into three sets instead of two. Predictive modeling in Enterprise Miner is iterative. The initial result is compared to the validation set and adjustments are made to the model. Once all adjustments are completed, the final testing is used to validate the results.

Figure 7. ROC Curves



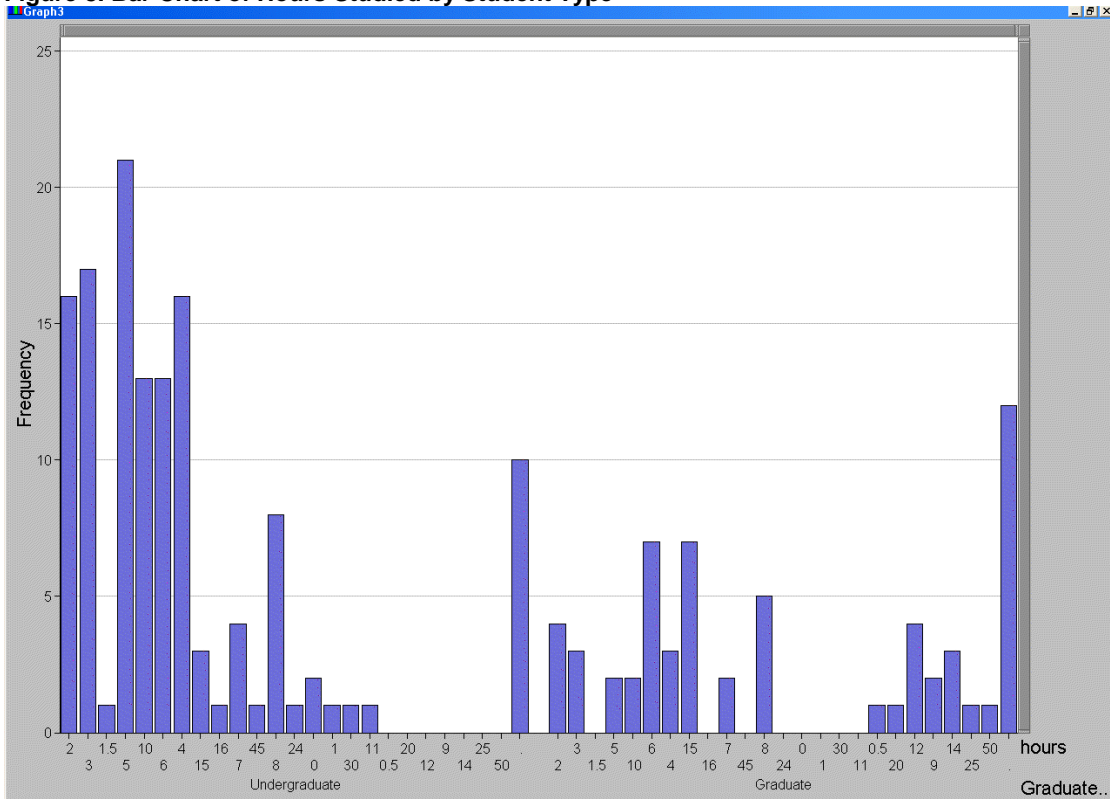
According to this measure, logistic regression gives the optimal outcome. However, the misclassification rate on the testing sample indicates that the decision tree gives a better overall result.

DATA VISUALIZATION

With large datasets, data visualization becomes an important part of exploring in data mining. Version 4.3 of Enterprise Miner included a node for SAS/Insight, and included all graphics within Insight. Version 5.1 removed the SAS/Insight Node, adding a Stat/Explore Node. However, neither yet provides the point-and-click graphics that are readily available in SAS Enterprise Guide.

There is, however, one set of graphics available in SAS/Stat that are not available in any other component of SAS. That procedure, PROC KDE, allows the investigator to overlay smoothed histograms to examine data. It is available for interval data only. In addition to questions concerning preferences for mathematics, students were asked to estimate the number of hours per week they expected to study for their mathematics courses. In Enterprise Miner, it is possible to use histograms to examine the data. Figure 8 was provided in Enterprise Miner using the StatExplore Node. Figure 9 shows the corresponding smoothed histogram as provided by SAS/Stat.

Figure 8. Bar Chart of Hours Studied by Student Type

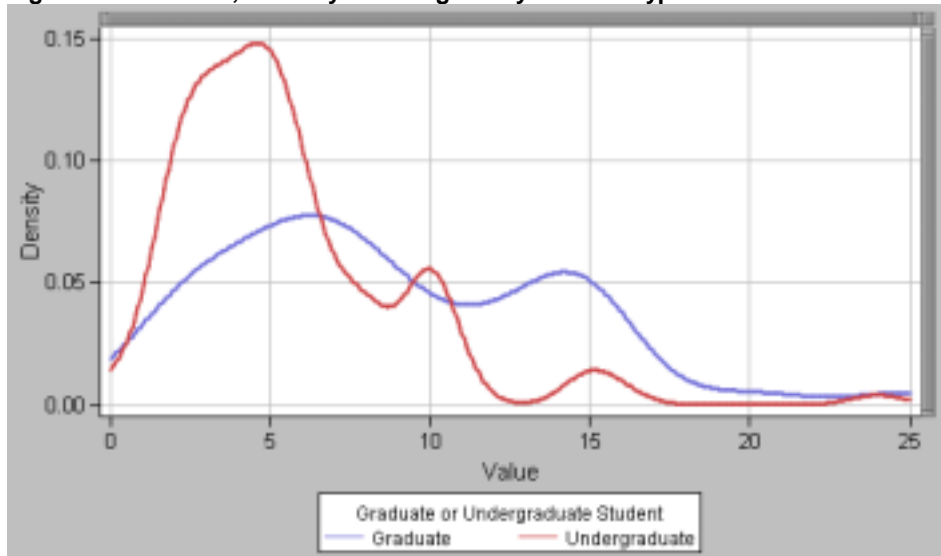


Comparing one to the other is difficult because grouped histograms are side-by-side. PROC KDE allows an overlay comparison:

```
Proc sort data=sasuser.studentsurvey;  
By student_type;  
Proc KDE data=sasuser.studentsurvey;  
Univar hours/gridl=0 gridu=25 out=sasuser.kdesurvey;  
  
By student_type;  
Run;
```

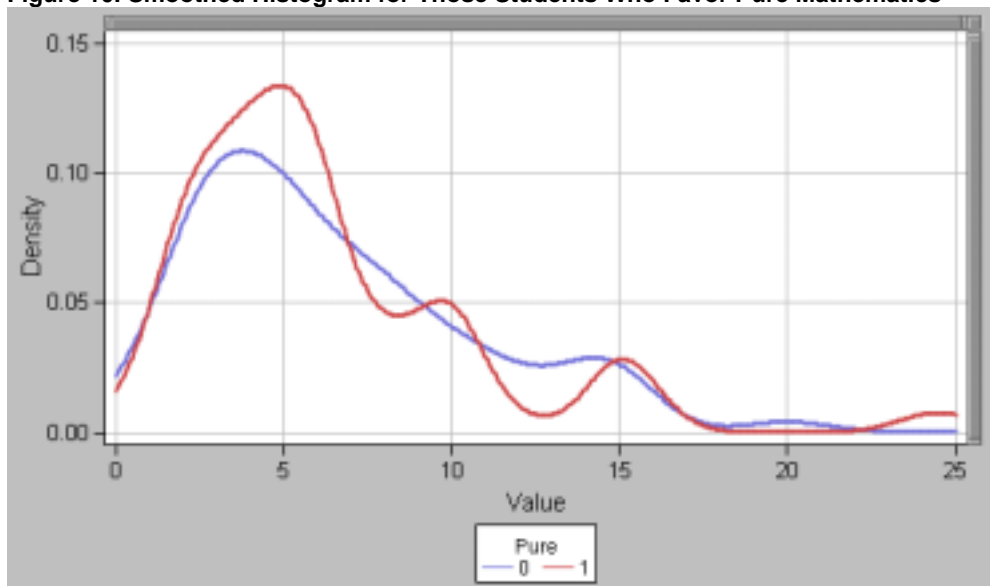
Proc GPlot is used to graph the results as given in Figure 9. Although Proc KDE has added graphics in version 9, the overlay still must be done with GPlot.

Figure 9. Smoothed, Over-layed Histogram by Student Type



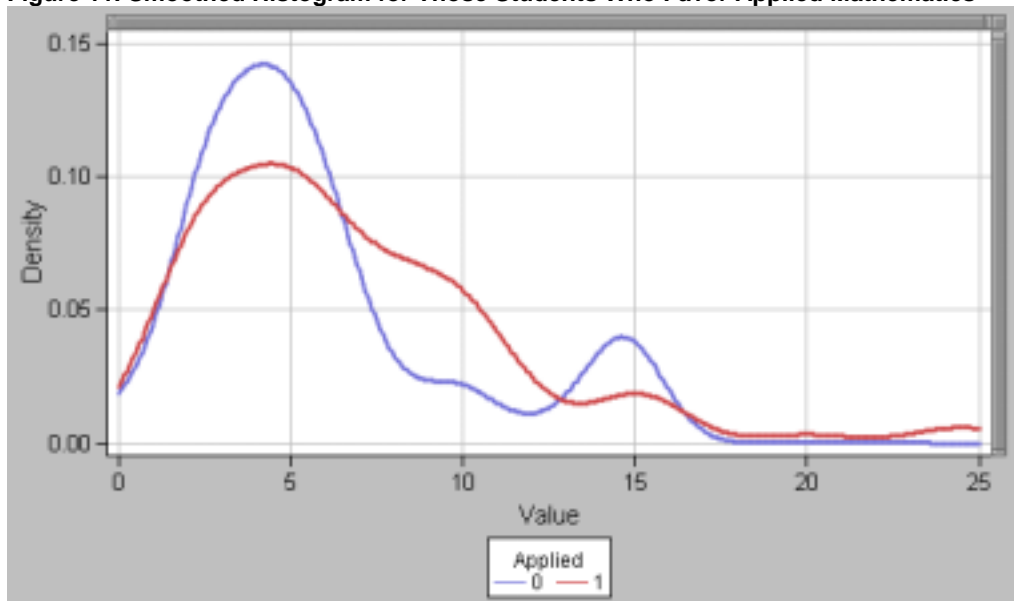
With the overlay, it is clear that almost all undergraduate students spend 10 hours or less in study compared to graduate students who are divided between 0-10 hours and 10-18 hours per week. In a similar comparison given in Figure 10, students who enjoy pure mathematics spend about the same amount of time in study as do the students who do not enjoy pure mathematics.

Figure 10. Smoothed Histogram for Those Students Who Favor Pure Mathematics



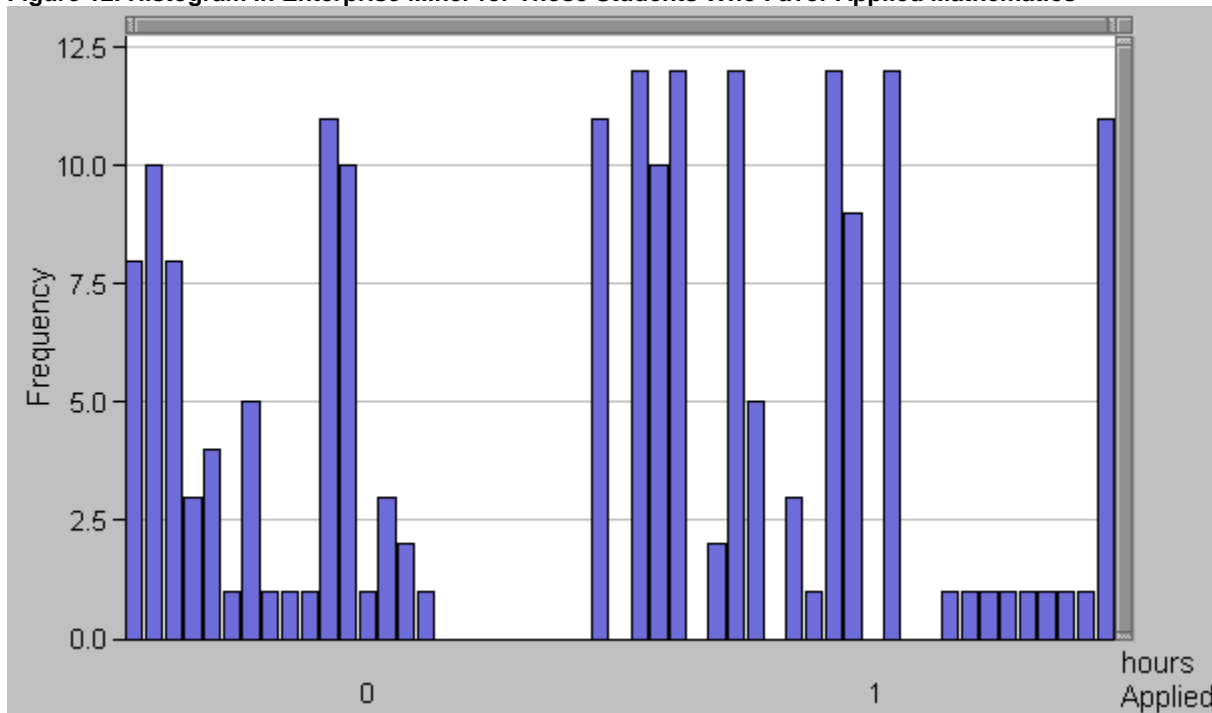
However, with the exception of a small group of students at the 15 hour mark who do not like pure mathematics, students who enjoy applied mathematics are more likely to study more than 7 hours per week.

Figure 11. Smoothed Histogram for Those Students Who Favor Applied Mathematics



The comparison is not as clear in the histogram provided in Enterprise Miner.

Figure 12. Histogram in Enterprise Miner for Those Students Who Favor Applied Mathematics



CONCLUSION

There are many similar procedures in SAS/Stat and Enterprise Miner that can be used for similar needs. However, generally the process is different since data are routinely partitioned to validate models, and to optimize the choice of model. Although SAS/Stat contains many similar models, it does not have the partitioning process readily available, although it can be coded into the process. There are some exploratory techniques built into SAS/Stat, such as PROC KDE, that can complement Enterprise Miner. In addition, partitioning and imputation steps can be performed in Enterprise Miner, and the data analyzed using SAS/Stat techniques.

CONTACT INFORMATION

Patricia Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6826
502-852-7132 (fax)
pcerrito@louisville.edu

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.

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