Paper 247-27

Discriminant Analysis, A Powerful Classification Technique in Data Mining

George C. J. Fernandez
Department of Applied Economics and Statistics / 204
University of Nevada - Reno
Reno NV 89557

ABSTRACT

Data mining is a collection of analytical techniques used to uncover new trends and patterns in massive databases. These data mining techniques stress visualization to thoroughly study the structure of data and to check the validity of the statistical model fit which leads to proactive decision making. Discriminant analysis is one of the data mining techniques used to discriminate a single classification variable using multiple attributes. Discriminant analysis also assigns observations to one of the pre-defined groups based on the knowledge of the multi-attributes. When the distribution within each group is multivariate normal, a parametric method can be used to develop a discriminant function using a generalized squared distance measure. The classification criterion is derived based on either the individual within-group covariance matrices or the pooled covariance matrix that also takes into account the prior probabilities of the classes. Non-parametric discriminant methods are based on non-parametric group-specific probability densities. Either a kernel or the k-nearest-neighbor method can be used to generate a non-parametric density estimate in each group and to produce a classification criterion. The performance of a discriminant criterion could be evaluated by estimating probabilities of mis-classification of new observations in the validation data. A user-friendly SAS application utilizing SAS macro to perform discriminant analysis is presented here. Car93 data containing multi-attributes is used to demonstrate the features of discriminant analysis in discriminating the three price groups, “LOW”, “MOD”, and “HIGH” groups.

INTRODUCTION

Data mining is the process of selecting, exploring, and modeling large amounts of data to uncover new trends and patterns in massive databases. These analyses lead to proactive decision making and knowledge discovery in large databases by stressing data exploration to thoroughly study the structure of data and to check the validity of statistical models that fit.

Discriminant Analyis (DA), a multivariate statistical technique is commonly used to build a predictive / descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups. In DA multiple quantitative attributes are used to discriminate single classification variable. DA is different from the cluster analysis because prior knowledge of the classes, usually in the form of a sample from each class is required. The common objectives of DA are i) to investigate differences between groups ii) to discriminate groups effectively; iii) to identify important discriminating variables; iv) to perform hypothesis testing on the differences between the expected groupings; and v) to classify new observations into pre-existing groups.

Stepwise, canonical and discriminant function analyses are commonly used DA techniques available in the SAS systems STAT module [SAS Inst. Inc. 1999]. CAR93 data containing multi-attributes, number of cylinders (X2), HP (X4), car width (X11), and car weight (X15) are used here to demonstrate the features of discriminant analysis in classifying three, “LOW (2)”, “MOD (3) ”, and “HIGH (1)” price groups. A user-friendly SAS macro developed by the author utilizes the latest capabilities of SAS systems to perform stepwise, canonical and discriminant function analysis with data exploration is presented here. The users can perform the discriminant analysis using their data by following the instructions given in the appendix and by downloading the SAS macro-call file from the author’s home page at http://www.ag.unr.edu/gf.

Data exploration and checking for multivariate normality and outliers.

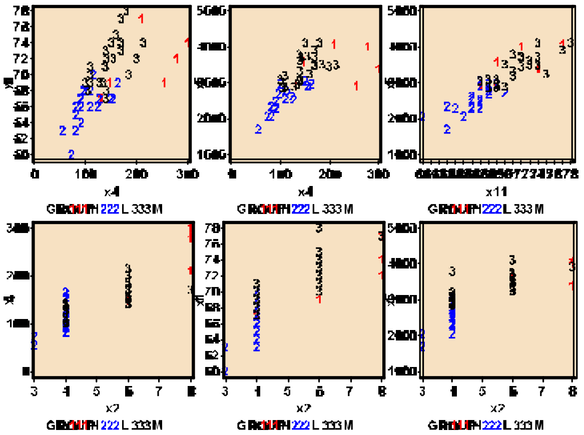


Figure 1 Group discrimination in a simple scatter plots

Examining the group discrimination based on simple scatter plots between any two discrimination variables is the first step in data exploring. An example of simple two-dimensional scatter plots showing the discrimination of three price groups is presented in Figure 1.

These scatter plots are useful in examining the range of variation and the degree of linear associations between any two attributes. The scatter plot presented in Fig.1 revealed the strong correlation existed between HP (X4) and weight (X15) and between HP (X4) and width (X11). These two attributes appeared to discriminate the three price groups to a certain degree.

Checking for multivariate normality: The right choice of selecting parametric vs. non-parametric discriminant analysis is dependent on the assumption of multivariate normality within each group. The car price data within each price group is assumed to have a multivariate normal distribution with a common covariance matrix. This multivariate normality assumption can be checked by estimating multivariate skewness, kurtosis, and testing for their significance levels. The Quantile-Quantile (Q-Q plot) plot of expected and observed distributions [Khattree and Naik 1995] of multi-attribute residuals after adjusting for the group means could be used to graphically examine for multivariate normality. The estimated multivariate skewness (44.934, p-val:0.001) and multivariate kurtosis (33.368, p-val : 0.0001) clearly supported the hypothesis that after adjusting for the group differences that these

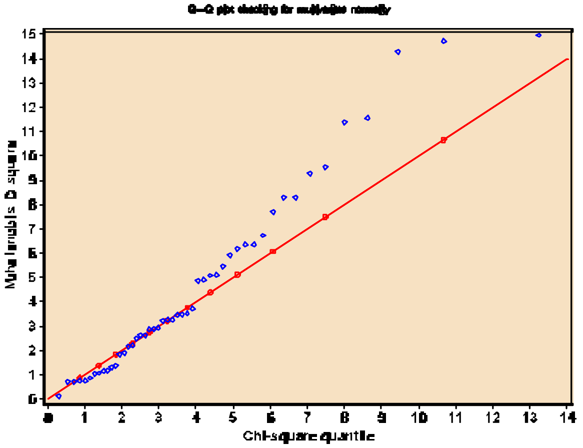


Figure2 Checking for Multi-variate normality in Q-Q Plot.

four multi-attributes do not have a joint multivariate normal distribution. A strong departure from the 45^0 angle reference line in the Q-Q plot (Fig. 2) supported this finding. Thus, DA based on non-parametric distributions could be considered as an appropriate technique for discriminating the three price groups based on the 4 attributes (x2, x4, x11 and x15).

Checking for multi-variate outliers Multivariate outliers can be detected in a plot between the difference of robust Mahalanobis distance - Chi-squared quantile vs. Chi-squared quantile value [Khattree and Naik 1995]. Observations 56, 82, 73, and 9 were identified as influential

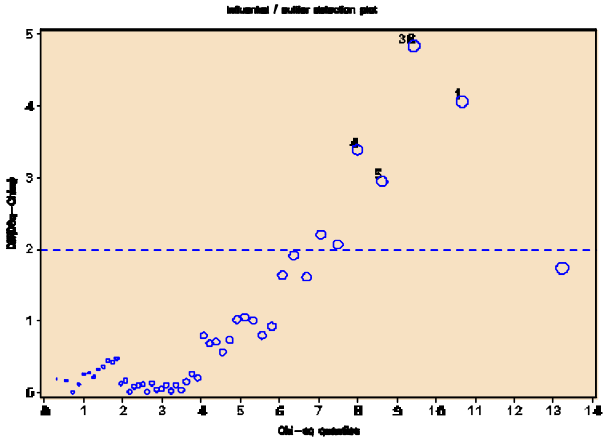


Figure 3 Checking for multi-variate outliers

observations since the differences between robust Mahalanobis distance and Chi-squared quantile value were larger than 5 and fall outside from the majority of the observations. It is important to examine the impact of these influential observations on discriminant analysis.

Canonical Discriminant Analysis (CDA): Canonical DA is a dimension-reduction technique similar to principal component analysis. The main objective of CDA is to extract a set of linear combinations of the quantitative variables that best reveal the differences among the groups. Given a nominal group variable and several quantitative attributes, the CDA extracts linear combinations of the quantitative variables (canonical variables) that capture between-class variation in much the same way that principal components summarize total variation [SAS Inst. Inc. 199]. Moreover, these canonical functions will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap. It is customary to standardize the multi attributes so that the canonical variables have means that are equal to zero and pooled within-class variances that are equal to one.

The extracted canonical variables have the highest possible multiple correlation with the groups. This maximal multiple correlation is called the *first canonical correlation*. The coefficients of the linear combination are the *canonical coefficients* or *canonical weights*. The derived composite variable defined by the linear combination is the *first canonical variable* or *canonical component*. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. In CDA, the process of extracting canonical variables are repeated until you extract the maximum number of canonical variable which is equal to the number of groups minus one, or the number of variables in the analysis, whichever is smaller. The standardized discriminant function coefficients indicate the partial contribution of each variable to the discriminant function(s), controlling for other attributes entered in the equation. The structure coefficients indicate the simple correlations between the variables and the discriminant function or functions. The structure coefficients should be used to assign meaningful labels to the discriminant functions. The standardized discriminant function coefficients should be used to assess each variable's unique contribution to the discriminant function. These structure loadings are commonly used when interpreting the meaning of the canonical variable because the structure loadings appear to be more stable, and they allow for the interpretation of canonical variable in the manner that is analogous to factor analysis.

Total Canonical Structure

Variable	Label	Can1	Can2
X2	cyl	0.826657	-0.390158
X4	hp	0.788772	-0.485631
X11	width	0.905626	0.158925
X15	weight	0.99717	0.031244

Table 1 Canonical structure loadings

The structure loadings for the first two canonical variables are presented in Table 1. The first and the second canonical functions accounts for 37% and 17% of the variation in the discriminating variables. All four multi attributes had the positive largest loadings on CAN1 and the X4 and X2 had a moderate negative loadings on CAN2.

The extracted canonical variable scores can be used to plot pairs of canonical variables in a two-dimensional bi-plots to aid visual interpretation of group differences. Inter-relationships among the four multi-attributes and the discriminations of the three groups are presented in Figure 4 & 5. The first canonical function which has largest loadings on all four attributes discriminated the LOW (2) price group from other two MOD (3) and HIGH (1) priced groups successfully. The second canonical function which has a moderate size loadings on X2and X4 discriminated the HIGH (1) price group from other two MOD (3) and LOW (2) priced groups successfully. Very small amount of overlap between the LOW(2) and MOD (3) groups were observed (Fig. 4).

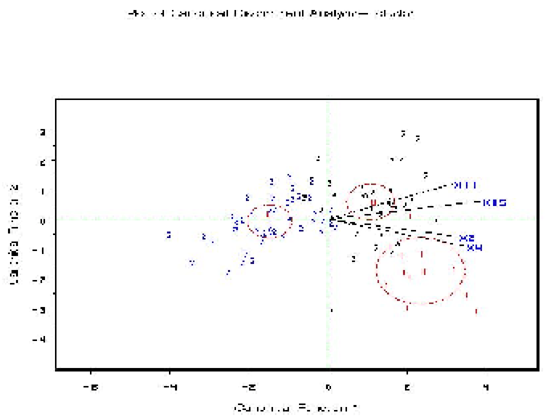


Figure4: Bi-plot display of multi-attributes and the group discrimination

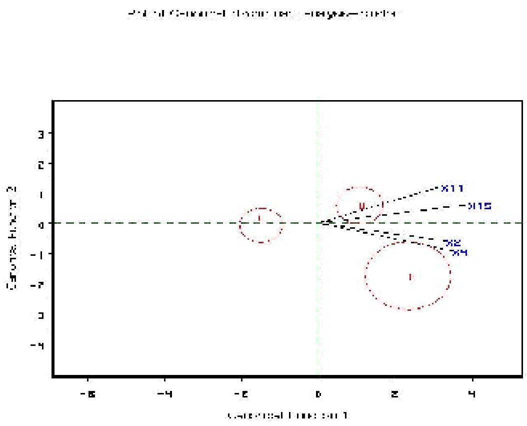


Figure 5: Biplot display of multi-attributes and the group means.

Predictive Discriminant Analysis (PDA): PDA is a predictive classification technique deals with a set of multi-attributes and one classification variable, the latter being a grouping variable with two or more levels. Predictive discriminant analysis is similar to multiple regression analysis except that PDA is used when the criterion variable is categorical and nominally scaled. As in multiple regression, in PDA a set of rules is formulated which consists of as many linear combinations of predictors as there are categories, or groups. A PDA is commonly used for classifying observations to pre-defined groups based on knowledge of the quantitative attributes. When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function using a measure of generalized squared distance. The discriminant function, also known as a classification criterion, is estimated by measuring generalized squared distance [SAS Inst. Inc. 1999]. The classification criterion can be derived based on either the individual within-group covariance matrices (a quadratic function) or the pooled covariance matrix (a linear function). This classification criterion also takes into account the prior probabilities of the discriminating groups. Each observation is classified in the group from which it has the smallest generalized squared distance. The posterior probability of an observation belonging to each class could be also estimated in PDA.

Classification results based on parametric Quadratic DF and error rates based on cross-validation are presented in Table 2. The overall discrimination was not satisfactory since none of the HIGH priced group was classified as HIGH and out of 25 medium priced cars, 7 of them were identified as LOW and 1 was HIGH. The overall error rate was 32% indicating about 1/3 of the cars were mis-classified based on parametric criterion. Therefore evaluation of other discriminating criterion becomes important.

Number of Observations and Percent Classified into GROUP

From GROUP	H	L	M	Total
H	0	1	4	5
	0	20	80	100
L	0	16	3	19
	0	84.21	15.79	100
M	1	7	17	25
	4	28	68	100
Total	1	24	24	49
	2.04	48.98	48.98	100
Priors	0.10204	0.38776	0.5102	

Error Count Estimates for GROUP

	H	L	M	Total
Rate	1	0.1579	0.32	0.3265
Priors	0.102	0.3878	0.5102	

Table 2: Classification results based on parametric Quadratic DF and error rates based on cross-validation

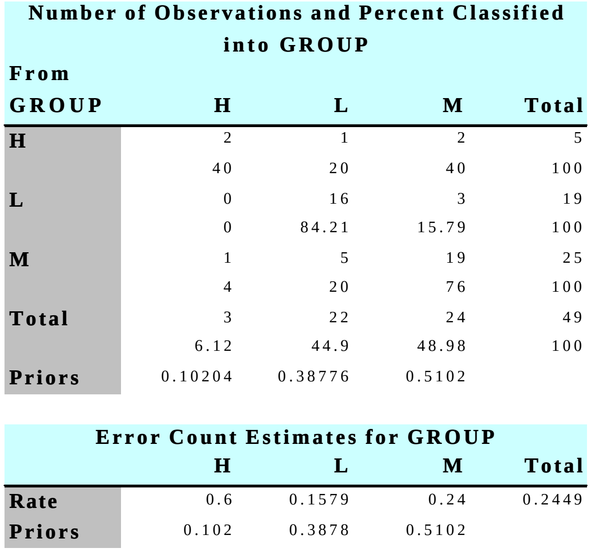
When no distribution assumptions within each group can be made, or when the distribution is not assumed to have multivariate normal, non-parametric methods can be used to estimate the group-specific densities. Non-parametric discriminant methods are based on nonparametric estimates of group-specific probability densities. Either a kernel method or the *k*-nearest-neighbor method can be used to generate a non-parametric density estimate in each group and to produce a classification criterion. The kernel method in SAS systems uses uniform, normal, Epanechnikov, biweight, or triweight kernels in the density estimation [SAS Inst. Inc. 1999]. Either Mahalanobis or euclidean distance can be used to determine proximity in the SAS DISCRIM procedure [SAS Inst. Inc. 1999]. When the *k*-nearest-neighbor method is used, the Mahalanobis distances are estimated based on the pooled covariance matrix. Whereas in the kernel method, the Mahalanobis distances based on either the individual within-group covariance matrices or the pooled covariance matrix is estimated.

In non-parametric DA estimation, with the estimated group-specific densities and their associated prior probabilities, the posterior probability estimates of group membership for each class can be evaluated. The classification of an observation vector **x** is based on the estimated group specific densities from the calibration or training sample. From these estimated densities, the posterior probabilities of group membership at **x** are evaluated.

The classification results based on 4th - nearest neighbor non-parametric PDA is presented in Table 3. All 5 observations were classified from HIGH to MOD & OTHER , 3 observations were classified from LOW to MOD and OTHER groups, and 4 observations were classified from MOD to LOW and OTHER groups. Thus an overall success rate of correct discrimination was about 76%. Classification results and error rates based on non-parametric K=4 nearest neighbor DFA was improved slightly compared with the parametric PDA.

The classification results based on non-parametric kernel density estimates with un-equal band width in Table 4. Three observation were classified from HIGH to LOW &MOD, 3 observations were classified from LOW to MOD, and 6 observations were classified from MOD to LOW & OTHER groups. Thus an overall success rate of correct discrimination was about 76%. The details of the mis-classified observations are presented in Table 4.

The performance of a discriminant criterion in the classification of new observations in the validation data could be evaluated by estimating the probabilities of mis-classification or error rates in the SAS DISCRIM procedure. These error-rate estimates include error-count estimates and posterior probability error-rate estimates. When the input data set is a SAS data set, the error rate can also be estimated by cross validation. SAS uses two types of error-rate estimates to evaluate the derived classification



Number of Observations and Percent Classified into GROUP

From GROUP	H	L	M	Total
H	2	1	2	5
	40	20	40	100
L	0	16	3	19
	0	84.21	15.79	100
M	1	5	19	25
	4	20	76	100
Total	3	22	24	49
	6.12	44.9	48.98	100
Priors	0.10204	0.38776	0.5102	

Error Count Estimates for GROUP

	H	L	M	Total
Rate	0.6	0.1579	0.24	0.2449
Priors	0.102	0.3878	0.5102	

Table 3: Summary of classification results and error rates based on non-parametric K=4 nearest neighbor DFA.

Number of Observations and Percent Classified into GROUP					
From GROUP	H	L	M	Other	Total
H	0	0	3	2	5
	0	0	60	40	100
L	0	16	2	1	19
	0	84.21	10.53	5.26	100
M	0	1	21	3	25
	0	4	84	12	100
Total	0	17	26	6	49
	0	34.69	53.06	12.24	100
Priors	0.10204	0.38776	0.5102		

Error Count Estimates for GROUP				
	H	L	M	Total
Rate	1	0.1579	0.16	0.2449
Priors	0.102	0.3878	0.5102	

Table 4: Summary of classification results and error rates based on non-parametric kernel density estimates with un-equal band width.

criterion based on parameters estimated by the training sample: i) error-count estimates and ii) posterior probability error-rate estimates. The error-count estimate is calculated by applying the discriminant criterion derived from the training sample to a test set and then counting the number of mis-classified observations. The group-specific error-count estimate is the proportion of mis-classified observations in the group. If the test sample set is independent of the training sample, the estimate is unbiased. However, it can have a large variance, especially if the test sample size is small [SAS Inst. Inc. 1999].

When no independent test sets are available, the same data set can be used both to calibrate and to evaluate the classification criterion. The resulting error-count estimate has an optimistic bias and is called an apparent error rate. To reduce the bias, the data can be split into two sets, one set for deriving the discriminant function and the other set for estimating the error rate. Such a split-sample method has the unfortunate effect of reducing the effective sample size.

Another way to reduce bias in estimating the classification error is cross validation [Lachenbrush and Mickey 1968]. In cross validation, *n*-1 out of *n* training observations in the calibration sample are treated as a training set. It determines the discriminant functions based on these *n*-1 observations and then applies them to classify the one observation left out. This is performed for each of the *n* training observations. The mis-classification rate for each group is the proportion of sample observations in that group that are mis-classified. This method achieves a nearly unbiased estimate but with a relatively large variance.

To reduce the variance in an error-count estimate Glick [1978] suggested a smoothed error-rate estimates. Instead of summing values that are either zero or one as in the error-count estimation, the smoothed estimator uses a continuum of values between zero and one in the terms that are summed. The resulting estimator has a smaller variance than the error-count estimate. The posterior probability error-rate estimates are smoothed error-rate estimates. The posterior probability estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate. When a parametric classification criterion (linear or quadratic discriminant function) is derived from a non normal population, the resulting posterior probability error-rate estimators may not be appropriate.

The smallest overall error rate based on cross-validation was 24.4% for the Car93 data based on the non-parametric kernel density estimates with un-equal band width (Table 4). The overall error rate is estimated through a weighted average of the individual group-specific error-rate estimates, where the prior probabilities are used as the weights. To reduce both the bias and the variance of the estimator, Hora and Wilcox [1982] compute the posterior probability estimates based on cross validation. The resulting estimates are intended to have both low variance from using the posterior probability estimate and low bias from cross validation. They use Monte Carlo studies on two-group multivariate normal distributions to compare the cross validation posterior probability estimates with three other estimators: the apparent error rate, cross validation estimator, and posterior probability estimator. They conclude that the cross validation posterior probability estimator has a lower mean squared error in their simulations.

The PDA based on non-parametric kernel density estimates with un-equal band width was selected as the best PDA analysis for this car93 data based on comparing the overall error rates for different PDA methods . The list of miss-classified observations and the posterior probability estimates are presented in table 5. The discriminant criterion derived from the training data set can be applied to a second independent validation validation data set

Posterior Probability of Membership in GROUP						
	From	Classified into				
Obs	GROUP	GROUP		H	L	M
2	H	L	*	0	0.6474	0.3526
4	H	M	*	0.0001	0	0.9999
5	H	M	*	0.0001	0	0.9999
10	L	M	*	0.001	0.3043	0.6947
11	L	M	*	0.0544	0.0459	0.8997
25	L	M	*	0.0011	0.0793	0.9196
37	M	L	*	0.0027	0.86	0.1372
41	M	L	*	0.0002	0.5704	0.4294
46	M	L	*	0.0099	0.81	0.18
47	M	L	*	0.0163	0.7927	0.191
48	M	H	*	0.7054	0	0.2946
50	M	L	*	0.0074	0.7135	0.2791

Table 5. Table of mis-classified observations based on non-parametric kernel density estimates with un-equal band width error rates based on cross-validation.

Number of Observations and Percent Classified into GROUP				
From				
GROUP	H	L	M	Total
H	7	0	2	9
	77.78	0	22.22	100
L	0	15	0	15
	0	100	0	100
M	4	3	12	19
	21.05	15.79	63.16	100
Total	11	18	14	43
	25.58	41.86	32.56	100
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for GROUP				
	H	L	M	Total
Rate	0.2222	0	0.3684	0.1969
Priors	0.3333	0.3333	0.3333	

Table 6. Classification results for an independent validation data set based on non-parametric kernel density estimates with un-equal band width error rates based on cross-validation

The classification results of the validation data set based on non-parametric kernel density estimates with un-equal band width is presented in Table 6. Two observation were classified from HIGH to MOD, and 7 observations were classified from MOD to LOW & HIGH groups. Thus an overall success rate of correct discrimination was about 81%.The classification results of the validation data set was equally comparable with the classification results of the training data set. However, due to the smaller sample size used in the training data set the applicability of the classification criterion derived has limited potential.

A user-friendly SAS application developed by the author utilizes the latest capabilities of SAS macros to perform stepwise, canonical and discriminant function analysis with data exploration is available now. The users can download the SAS macro-call file and perform the discriminant analysis reported in this paper using their data by following the instructions given in the Appendix and by downloading the SAS macro-call file from the author's home page at http://www.ag.unr.edu/gf.

REFERENCES

1. Glick N (1978) Additive estimators for probabilities of correct classification Pattern Recognition 10: 211-222[1].
2. Hora S. C and Wilcox J.B. (1982) Estimation of error rates in several population discriminant analysis. J. of. Marketing Research. 19:57-61
3. Khattree R. and Naik D.N , (1995) Applied Multivariate statistics with SAS software Cary NC . SAS Institute Inc.
4. Lachenbruch P. A and Mickey M.A. (1968) Estimation of error rates in discriminant analysis Technometrics 10 1-10
5. SAS Institute Inc. (1999) SAS/STAT Users Guide, Version 8, Cary NC .SAS Institute Inc

CONTACT ADDRESS

Dr. George C.J. Fernandez
Experimental Station Statistician and Associate Professor in Applied Statistics
Department of Applied Economics/Statistics/204
University of Nevada- Reno Reno NV 89557.
(775) 784-4206 E-mail: GCJF@unr.edu

AUTHOR'S BIO

Dr. George C.J. Fernandez currently serves as the statistician for the Nevada Experimental station and Cooperative Extension. He is also an Associate Professor in the department of Applied economics and statistics. He has more than 14 years of experience in teaching courses such as introductory statistical methods, design and analysis of experiments, linear and non-linear regression, multivaraite statistical methods and SAS programming. He is a professional SAS programmer and has over 20 years experience in SAS/BASE, SAS/IML, SAS/STAT , SAS/QC SAS/ETS, SAS/INSIGHT, SAS/ANLALYST, SAS/LAB, SAS/ASSIST and SAS/GRAPH. He has won best paper and poster presentation awards at the regional and international conferences. Recently, he has presented invited full-day workshops on “Applications of user-friendly statistical methods in Data mining at the American Statistical Association Joint meeting in Atlanta and at the Pre-Western SAS users Conference in Arizona. Many international and national SAS users are currently using his user-friendly SAS applications for data analysis via on-line. He has organized 7th Western Users of SAS conference (WUSS) at Los Angeles in 1999 and currently serving as the vice president for WUSS executive committee Inc..

APPENDIX

Instructions for downloading and running the DISCRIM SAS macro:

As an alternative to the point-and-click menu interface modules, a user-friendly SAS macro application to perform a complete discriminant analyis developed by the author is now available for the public to download . This macro approach integrates the statistical and graphical analysis tools available in SAS systems and provides complete data analysis tasks quickly without writing SAS program statements by running the SAS macros in the background. The main feature of this approach is that the users can perform graphical discriminant analysis quickly using the SAS macro-call file available for downloading. Using this MACRO APPROACH, the analysts can effectively and quickly perform complete data analysis and spend more time in exploring data, interpretation of graphs and output rather than debugging their program errors etc.

REQUIREMENTS:

This macro application was developed in Win/NT SAS version 6:12 and was tested in both version 6:12, 8.1, and 8.2. The requirements for using these SAS MACROs are:

1) A valid license to run the SAS software on your PC.

2) SAS modules such as SAS/BASE, SAS/STAT, SAS/IML and SAS/GRAPH should be installed in your computer to get the complete results.

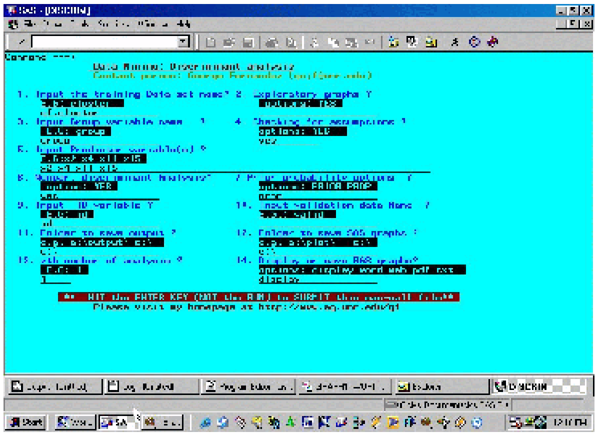
3) A working internet connection to access the DISCRIM macro while executing the the downloaded macro-call file.

The steps for performing discriminant analysis by the user-friendly SAS MACRO “DISCRIM”:

Step 1: Create a SAS data set
This data should contain the following variables: One classification group (GROUP) variable, which should be a categorical variable. and several continuous or numeric multi attribute variables.

Step 2: Downloading the macro-call file:
Visit the home page at http://www.ag.unr.edu/gf, click the running puppy dog and follow the instructions given in the page, to go to the SAS macro download page. Go to the miscellaneous section and download the DISCRIM.SAS MACRO-CALL file by clicking the sample demo link, and save this file to a disk and open it in the SAS program editor window. Click the RUN icon to open the DISCRIM MACRO-CALL window (Figure 6).

Step 3: Input the required values by following the instructions provided in the SAS MACRO-CALL window, DISCRIM (Figure 6).



```
Command ===>
            Data Mining: Discriminant analysis
            Contact person: George Fernandez (gcj@unr.edu)

 1. Input the training Data set name? 2. Exploratory graphs ?
    [illegible]                          options: YES
    [illegible]
 3. Input Group variable name ?       4. Checking for assumptions ?
    E.G.: group                          options: YES
    group                                yes
 5. Input Predictor variable(s) ?
    E.G.: x4 x11 x15
    x2 x4 x11 x15
 6. Stepwise discriminant Analysis?   7. Prior probability options ?
    options: YES                         options: PRIOR PROP
    yes                                  prop
 9. Input ID variable ?               10. Input validation data Name ?
    E.G.: id                             E.G.: valid
    id
11. Folder to save output ?           12. Folder to save SAS graphs ?
    e.g. c:\output\ c:\                  e.g. c:\plots\ c:\
    c:\                                  c:\
13. zth number of analysis ?          14. Display or save SAS graphs?
    E.G.: 1                              options: display word web pdf txt
    1                                    display

   ** HIT the ENTER KEY (NOT the RUN) to SUBMIT the macro-call fields**
            Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 6. Screen shot of the DISCRIM macro-call window

Macro–call options:

1. Input the name of the temporary or permanent SAS data set

2. To perform Exploratory analysis and stepwise discriminant analysis type YES in the field #2. No output for CDA and PDA are not produced.

3. Input the discriminating GROUP variable name.

4. To check for multivariate normality and outliers type YES in the field 4.

5. List all the continues multi-attribute variables names.

6. Input prior probability option

7. To perform non-parametric discriminant analysis, type YES .

8. Input any optional ID variable name

9. Input the name of the Optional VALID data set.

10. Input the folder name to save your SAS output (HTML, RTF, PDF, TXT) files.

11. Input the folder name to save your SAS graphic files.

12. Input the z th number of analysis. Change the value of Z by an increment of one when you repeat the analysis. This avoids over-writing the saved output file.

13. Display or save the graphics and SAS out put files

Options for saving the SAS output and SAS graphics files.

Users can select the folders to save the SAS output (Text file in SAS Version < 8.0, and RTF file and HTML files in version 8.1 and above, PDF file in version 8.2 and above) and the graphics files by inputting the folder names in the MACRO-CALL window. Also, the users can select one of the following graphic file format when saving the graphics produced by the SAS MACRO:

Display: Files are not saved, but displayed in the SAS graphics Window.
TXT: CGM files suitable for including in Corel Word perfect and WORD products.
RTF: JPEG files suitable for including in power point presentations .

Step 4: Submit the SAS MACRO:

After inputting all required fields, move your cursor to the last MACRO field, and hit the ENTER key to run the SAS MACRO. (Do not click the RUN icon). The MACRO-CALL window file, automatically accesses the DISCRIM SAS MACROs from the Internet server, College of Agriculture, University of Nevada and provide the users the required exploratory graphs, and the results of CDA and PDA analysis in Txt, RTF, HTML, and PDF format.