REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
31-12-2002	Final Report	1 June 2001 – 30 September 2002

4. TITLE AND SUBTITLE
Classification and Selection for Personnel Applications Using a Data Envelopment Analysis Approach

5a. CONTRACT NUMBER
N00014-01-1-0917

5b. GRANT NUMBER
N00014-01-1-0917

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Retzlaff-Roberts, Donna – The University of Memphis
Dula, Jose H., The University of Mississippi
Van Scotter, J., The University of Memphis

5d. PROJECT NUMBER
3712

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Memphis, Fogelman College of Business, Memphis, TN 38152
The University of Mississippi, School of Business, University, MS 38677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project develops and tests a new approach for improving the effectiveness of personnel selection and classification decisions. There are three main data characteristics that cause difficulty for existing classification methods (1) unbalanced group sizes, (2) unequal misclassification costs, and (3) non-normal data. Preliminary research suggests a hybrid method incorporating data envelopment analysis (DEA) and linear programming discriminant analysis (DEA/DA) is effective in this difficult situation and outperforms other methods. Research is conducted to fully develop and test this promising methodology. Results suggest ways DEA/DA may help alleviate long-standing selection and classification problems.

15. SUBJECT TERMS
data envelopment analysis, discriminant analysis, personnel classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	UU	5	Tanja F. Blackstone
u	u	u			19b. TELEPHONE NUMBER *(Include area code)* 901-874-4642

FINAL TECHNICAL REPORT

GRANT #: N00014-01-1-0917

PRINCIPLE INVESTIGATOR: Donna Retzlaff-Roberts (email:retzlaff@memphis.edu)

CO-INVESTIGATORS: Jose H. Dula (email: jdula@olemiss.edu)
James Van Scotter (email: jvanscot@memphis.edu)

INSTITUTIONS: The University of Memphis (Retzlaff-Roberts and Van Scotter)
The University of Mississippi (Dula)

GRANT TITLE: Classification and Selection for Personnel Applications Using A Data Envelopment Analysis Approach

AWARD PERIOD: 1 June 2001 – 30 September 2002

OBJECTIVE: The objective of this study is to develop and test the Data Envelopment Analysis method of discriminant analysis (DEA/DA) to determine under what circumstances it is most effective and how it compares to existing methods for data sets that have the challenging characteristics of unbalanced group sizes and are not multivariate normal, as is frequently the case for personnel data.

APPROACH: Data sets with varying degrees of (1) unbalance in group size, (2) unequal misclassification costs, and (3) departures from multivariate normality are used. These data sets are utilized in developing and testing the DA/DEA method with various possible model formulations being studied. The DA/DEA results are compared with those obtained using other discriminant methods, and the accuracy and effectiveness of each method evaluated for each condition.

ACCOMPLISHMENTS (throughout award period):

We have completed the study of the data envelopment analysis approach to classification. This involves a detailed analysis of methods including four different formulations of linear programming and statistical discriminant analysis. The four linear programming methods differ on the form of the normalization constraint used and also on how the variable weights are constrained. The models are:

1. DEA Ratio: The Ratio model seeks to minimize the ratio of eternal deviations to internal deviations. This method has the advantage of avoiding the distortion problem where the data is "elongated" in some directions while being compressed in others. This distortion problem interferes with the comparison of alternate solutions. There are two variations of this model, one where the variables are divided into two categories traditionally called inputs and output in DEA. Here, in more general manner the variables that are better high form one group, which is usually called the outputs but will be called the positives here. The other variables, that are better low, form the other groups which are usually called inputs but will be called negative here. The weight for each variable is constrained as in DEA to be either positive or negative.

2. Ratio: In the second variation of the ratio model the variables are not assigned to groups a priori and the weights are unrestricted to assume either positive or negative values.

3. DEA: The third formulation uses a DEA type normalization constraint where the weighted sum of positives summed across all cases must equal some constant. Variable weights are also constrained to be either positive or negative based on whether each variable is a positive or a negative.

4. Midpoints: A fourth formulation was developed to use linear programming model to maximize the difference in the midpoints of the two groups. This is the principle upon which statistical discriminant analysis is based and assumes that the univariate scores form a normal distribution for each group. Thus maximizing the difference in the group averages will not necessarily minimize the amount of overlap between groups.

All of these models were tested via application to the same data sets. Data sets varied based on:

1. The degree of unbalance in the groups. This was measured based on the percent of cases in the failing group. This was varied from as low as 3% to as high as 30%.

2. The degree of non-normality in the data set. This was controlled based on the number of variables involved as well as the ratio of scale variables to categorical variables. Many of our variables were binary nature with a few scale variables.

We developed a two-stage method for discriminant analysis whereby the first stage is to obtain the discriminant function. Using this function each unit receives a discriminant score and reduces the data from multivariate to univariate. The second step is to determine where the "threshold" or cut-off point should be located to divide the cases into two groups. This is done using the two types of misclassification costs such that total misclassification cost is minimized. This two-stage approach allow for adjusting the overall rejection rate. This is important for unbalance groups because the more unbalanced a data set is the better classification results will be for low rejection rates.

CONCLUSIONS: The two ratio methods consistently outperformed the DEA method and the midpoint method. This is an interesting result because the two ratio methods are the ones that avoid the data distortion problem while the other two methods do not. Thus it appears that the avoidance of data distortion is beneficial in obtaining the best classification solution. The midpoint method may also be inferior is due to the fact the assumption of normality for the univariate scores is violated when the variables are non-normal. In fact, it was found that the distribution of scores for the smaller failing group tend to be much more skewed that those of the larger passing group. When score distributions are skewed, maximizing the distance between the group averages may not produce the best solution.

Between the two ratio methods, the unconstrained ratio method sometimes outperformed the DEA ratio method. These methods differ based on whether or not the variable weights are constrained a priori based on knowledge of whether a particular variable is a positive or a negative. This relates to the two different "schools of thought" regarding DEA and previous models of linear programming discriminant analysis. Within the field of DEA it is commonly believed that it is beneficial to take advantage of prior knowledge concerning the effect of each variable, and this is certainly intuitive. Much of the linear programming discriminant analysis literature, on the other hand, deems complete mathematical freedom desirable.

When the constrained ratio model outperformed the DEA ratio model there were one or more variables receiving weights with the opposite sign from what was expected. This behavior is certainly similar to multicollinarity found in regression. Linear programming is non-parametric, but correlations among variables may result in weights taking on the opposite of the expected sign, which is counterintuitive. DEA's use of constrained weights results in weights being equal to zero when this situation occurs and prevents the counterintuitive result of an input with a negative weight, or an output with a positive weight.

Comparing the linear programming methods to statistical discriminant analysis, linear programming is better suited as the percent of unbalance in the groups increases. As degree of non-normality in the data increases the linear programming methods tend to outperform statistical discriminant analysis.

It was found that at times an excessive amount of weight was placed on one variable. This was measured based on what percent of the total score for the cases came from a particular variable. For example, if on average one particular variable accounts for 70% of the scores' values, then an excessive amount of weight is placed on one particular variable. How high this percentage can go is of course a subjective decision. The use of bounds in the range of 33% to 50% was studied, and placing an upper bound on the magnitude of weights was found to be desirable.

The two-stage approach is recommended when differing costs of the two types of misclassification is of concern. This two-stage approach finds the discriminant scores in stage one, then sets the "threshold" or cut-off value in stage two. The threshold is chosen to minimize the total cost of misclassification. When misclassifying into the successful group is higher cost, then the threshold is raised to classify more cases into the failing group. This situation would occur when there is an adequate supply of candidates or applicants to a program and the training cost is high, thus the priority is to admit only those with a very high probability of success. Conversely, when the cost of misclassifying into the failing group is higher, then the threshold will be lowered to classify more units into the successful group. This situation would occur when candidates/applicants are in short supply, thus there is a high opportunity cost of rejecting a candidate who would have been successful.

In final conclusion we recommend that linear programming discriminant analysis be considered when the small group contains less than 10% of the cases and the data is comprised primarily of non-normal variables such as binary or ordinal variables. The particular model formulation we recommend is the DEA ratio formulation. This model avoids the data distortion problem and constrains the variable weights based on prior knowledge of whether each variable if better higher or lower, as is the norm in Data Envelopment Analysis. Personnel data frequently has these characteristics of unbalance and non-normal data, thus this linear programming approach offers numerous applications.

Possible extensions include enhanced classification ability for a variety of applications both within and beyond personnel data. Within personnel data, examples of other applications include prediction of likelihood to reenlist, prediction of likelihood to complete a term of service, etc. The need to accurately predict binary outcomes such as yes/no or success/failure is pervasive to a wide variety of applications.

SIGNIFICANCE: The significance of this work is improved ability for classification of personnel for the difficult situation of highly unbalanced groups and non-normal data. This includes a reduction in both type I and type II error such that the right personnel can be selected for programs to maximize the probability of success and minimize the probability of failure.

PATENT INFORMATION:

None

AWARD INFORMATION:

None

REFEREED PUBLICATIONS (for total award period):

None as yet.

BOOK CHAPTERS, SUBMISSIONS, ABSTRACTS AND OTHER PUBLICATIONS (for total award period)

1. Retzlaff-Roberts, D., Dula, J.H., Van Scotter, J., "Linear Programming Methods of Discriminant Analysis for Unbalanced and Non-Normal Data," working paper in progress.