DETERMINING THE EFFICACY OF MATHEMATICAL PROGRAMMING APPROACHES FOR MULTI-GROUP CLASSIFICATION

by

DINESH R. PAI

A Dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Dr. Kenneth D. Lawrence

and approved by

_____

_____

_____

_____

Newark, New Jersey

October, 2009

ABSTRACT OF THE DISSERTATION

DETERMINING THE EFFICACY OF MATHEMATICAL PROGRAMMING APPROACHES FOR MULTI-GROUP CLASSIFICATION

By Dinesh R. Pai

Thesis Director: Dr. Kenneth D. Lawrence

Managers have been grappling with the problem of extracting patterns out of the vast database generated by their systems. The advent of powerful information systems in organizations and the consequent agglomeration of vast pool of data since the mid-1980s have created renewed interest in the usefulness of discriminant analysis (DA). Expert systems have come to the aid of managers in their day-to-day decision making with many successful applications in financial planning, sales management, and other areas of business operations (Erenguc and Koehler 1990).

Currently, no comprehensive research study exists that tests the robustness of multi-group classification analysis. Our research aims to bridge the gaps in the existing works and take a step further by extending our study to four-group classification problems. The main purpose of this research is to determine the efficacy of mathematical programming classification models, more specifically, LP methods vis-à-vis statistical approaches such as discriminant analysis (Mahalanobis) and logistic regression, an artificial intelligence (AI) technique such as a neural network, and a non-parametric technique such as k-nearest neighborhood (k-NN) for four-group classification problems. This research also

proposes an integrated (hybrid) model that combines a non-parametric classification technique and a LP approach to enhance the overall classification performance. Furthermore, the study extends an existing two-group LP model (Bal et al. 2006) based on the work of (Lam and Moy 1996b) and apply it to four-group classification problems. These models are tested through robust computational experiments under varying data conditions using a financial product example. The characteristics of a real dataset are used to simulate (Monte Carlo method) multiple sample runs for four group classification problems with three continuous independent variables.

The experimental results show that LP approaches in general and the proposed integrated method in particular consistently have lower misclassification rates for most data characteristics. Furthermore, the integrated method utilizes the strengths of both the methods: k-NN and linear programming, thereby considerably improving the classification accuracy.

ACKNOWLEDGEMENTS

The writing of this dissertation has been a significant academic challenge and an enriching experience. Without the support, patience, and guidance of many people, this study could not have been completed.

I express my deepest gratitude to my dissertation chair, Professor Kenneth D. Lawrence who went beyond the call of duty in ensuring the completion of this work. I have immensely benefitted from his advice and experience, on both academic and personal fronts. But for his humorous demeanor, it would not have been possible for me to circumvent the stress levels common with graduate students.

I take this opportunity to thank Professor Ronald D. Armstrong, whom I have known from day zero of the PhD program, for his patience and guidance throughout this program, and for acceding to my request to be on my committee. It would have been difficult to traverse this path without his understanding, help and generosity.

Professor Ronald K. Klimberg not only provided important advice regarding computational experiments, but more importantly, advice related to career planning. I have highly benefitted from the numerous erudite conversations I have had with him pertaining this work. Thank you.

Thanks are also due to Professor Sheila M. Lawrence for her encouragement and her ever willingness to help with editing draft papers and dissertation. I am confident my skill in this area is much better than it was at the start of this program.

Professor Lei Lei has been a tremendous source of inspiration and help throughout my stay at Rutgers. I would like to especially thank her and the Rutgers

Supply Chain Center for the generous support for procuring software's required for completing this study. She is indeed a class act.

The PhD program office has been of great help all these years. I would like to especially thank Goncalo Filipe for his 'infinite' patience and helpful nature. I would also like to thank my colleagues: Deepa Aravind, Amy Chen, Sungyong Choi, Adam Fleischhacker, Su Gao, Katie Martino, Karthik Puranam, Junmin Shi, Ari Yezegal and others, with whom I have had interesting discussions at various points in time, shared some light moments, and gained a lot from their experience and suggestions.

Special thanks to my wife, Ananya Pai, for her selfless dedicated support throughout and for providing us with 'pot of joy' our daughter Mythili Pai. Thanks are also due to my brothers and their family, and my friends for their constant support and encouragement throughout the program. Above all, I would like to thank and dedicate this work to my parents who have inculcated in me the importance of education early in my life and without whose myriad sacrifices and blessings I would not have accomplished this.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Research Motives

Managers have been grappling with the problem of extracting patterns out of the vast database generated by their systems. The advent of powerful information systems in organizations and the consequent agglomeration of vast pool of data since the mid-1980s have created renewed interest in the usefulness of discriminant analysis (DA). Expert systems have come to the aid of managers in their day-to-day decision making with many successful applications in financial planning, sales management, and other areas of business operations (Erenguc and Koehler 1990).

Much research has been done on the application of discriminant and classification techniques to *a priori* predictive as well as descriptive segmentation. In *a priori* predictive approaches, the type and number of segments are determined in advance based on a set of criteria, and subsequently, predictive models are used to describe the relationship between the segment membership and a set of independent variables (Wedel and Kamakura 1998). The main approaches for *a priori* descriptive segmentation are based on statistical or operations research methods of classification. More recently, a neural network technique has been used successfully in various classifications applications (Ripley 1994). The statistical methods include discriminant analysis, logistic regression, regression and cross-tabulation, while the operations research techniques include mathematical programming (MP) methods such as linear programming and its variants: goal programming and fuzzy goal programming (Thomas et al. 2006).

Most of the classical statistical techniques such as discriminant analysis (Mahanalobis), logistic regression, multiple regression, and others make strong parametric assumptions such as multivariate normal populations with same variance/covariance structure, absence of multicollinearity, and absence of specification errors (Meyers et al. 2006). However, many real-life data sets do not satisfy such underlying assumptions.

Since the early 1980s, considerable research has been devoted to mathematical programming (MP) methods, more specifically for two-group classification problems. These researches have highlighted several interesting characteristics. First, MP methods make no rigid assumptions about the functional form and hence are distribution free. This fits well with the real-life data sets which are invariably contaminated. Second, they do not require larger datasets and are less sensitive to outliers. Finally, MP methods require considerable computing time, but a continuous drop in computing cost and an increase in computing power has overcome this drawback and made these methods practical. The MP methods can be further classified into (a) linear programming approaches, (b) nonlinear approaches, and (c) MIP approaches. Meanwhile, a shortcoming of MP methods is that they not as amenable to statistical inferences as are statistical DA approaches (Sueyoshi and Hwang 2004).

However, most of the research has been focused on two-group classification problems across all techniques. In the past decade or so, there have been several researches aimed at three-group classification using MP approaches such as linear programming (LP) and

mixed integer programming (MIP) methods. Lam and Moy (1996b) propose a LP approach for three-group classification problem. In this approach, the authors aggregate information regarding weights, instead of computing cut-off scores and claim to provide better estimates of group boundaries. (Pavur and Loucopoulos 1995, Loucopoulos and Pavur 1997, Loucopoulos 2001) use MIP approach for three-group classification and develop several variants of their base MIP model. In all of the above multi-group studies the authors have used small to moderate size datasets. Moreover, performance of their models vis-à-vis other commonly used classification techniques such as logistic regression, or AI technique such as a neural network to test the robustness of their proposed models is yet to be determined. Since the models use small datasets, further research is needed to assess, the classification performance of these models with moderate to large datasets with different group configurations, data characteristics, and computation efficiency.

1.2 Research Objectives

Our research aims to bridge these gaps and take a step further by extending our study to four-group classification problems. The main purpose of this research study is to determine the efficacy of MP classification models, more specifically, LP methods vis-à-vis statistical approaches such as discriminant analysis (Mahalanobis), and logistic regression, an artificial intelligence (AI) technique such as a neural network, and a non-parametric technique such as k-nearest neighborhood (k-NN) for four-group classification problems. This study also proposes an integrated (hybrid) method that combines a non-parametric classification technique and a LP approach to enhance the

overall classification performance. Furthermore, the study extends an existing two-group LP model (Bal et al. 2006) based on the work of (Lam and Moy 1996b) and apply it to four-group classification problems. These models are tested through robust experimental design using a financial product example. Through the development and testing of multiple models this study aims to provide significant insight into many important questions concerning these classification approaches, such as the following:

1. Can a combination of a simple machine learning algorithm and a non-parametric approach yield better results compared with statistical and AI techniques?
2. Do LP approaches perform better in terms of lower misclassification rates than the statistical and AI techniques for multi-group classifications problems?
3. How do the models behave under different data and group characteristics?
4. Are there significant differences in the individual error rates for different classification approaches?

This study plans to contribute new knowledge through each of these important questions. In addition, to the best of our knowledge, this is the first comprehensive study which evaluates statistical, AI, and LP approaches for multi-group classifications using robust experimental design and a financial product example with moderate to large data sets.

Previous work pertaining LP approaches that contributes directly to the present study, such as Freed and Glover (1981a), Lam et al. (1996a), Kiang (2003), Bal et al. (2006), will be reviewed in detail, with respect to both its content and methodology, in Chapter 2.

The purpose of this introductory chapter is to discuss the questions examined in this study, foreshadowing what is to come. As such, it is necessary to provide some conceptual and historical perspective about discriminant and classification techniques. The research questions, which are the focus of this proposal, will follow from the subsequent discussion.

1.3 Historical Background

As in statistical regression, the objective of classification technique is to identify a functional relationship between a response (dependent) variable Y and a vector of explanatory (independent) variables or attributes X from a given set of observations (Y, X). However, in classification methods the response variable is discrete (dichotomous or polytochomous) where as in statistical regression it is real valued variable. The response variable is denoted by $C_1, C_2, ..., C_q$, where q is the number of pre-specified groups or classes (Doumpos and Zopounidis 2002). The objective of the classification methods is to first analyze the training data set and develop a model for each class or group using the attributes available in the data. Once the model developed using training set performs satisfactorily, it can be used to classify future independent test data.

In general, classification models assign observations of unknown class membership to a number of specified classes or groups using a set of explanatory variables associated with the group. These models have found myriad business applications such as in credit

evaluation systems (Myers and Forgy 1963), differentiating bank charge-card holders (Awh and Waters 1974), screening credit applicants (Capon 1982), assessing project implementation risk (Anderson and Narasimhan 1979), predicting consumer innovators for new product diffusion (Robertson and Kennedy 1968), predicting corporate bankruptcy (Altman 1968), investigating new product success or failure (Dillon et al. 1979), predicting bank failures (Tam and Kiang 1992), and approving loan applications (Gallant 1988). These models have been particularly useful in market segmentation based on observable and product specific bases. The advances in computers and information technology have further increased the efficacy of such approaches whereby vast amount of historical customer data can be processed to understand customer needs and wants. This has resulted in more focused marketing strategy resulting in lower costs, higher response rates, and consequently higher profits (Zahavi and Levin 1997b).

Since Fisher's seminal work (Fisher 1936) on linear discriminant analysis numerous methods have been developed for classification purposes. Discriminant analysis has been successfully applied in many business applications including building credit scoring models for predicting credit risk, and investigating product failures (Dillon et al. 1979, Myers and Forgy 1963). Logistic regression is a related statistical method which is now widely used and (Westin 1973) was one of the first to apply it in a binary choice situation. Mangasarian (1965) was the first to use LP method in classification problems for distinguishing between the elements of two disjoint sets of patterns. Freed and Glover (1981a) extended this work for predicting the performance of job applicants based

on a set of explanatory variables. Tam and Kiang (1992) were one of the first to use a neural network in business research for predicting bank failures.

The two-group classification problems have been extensively dealt with in the literature, till date. Srinivasan and Kim (1987), Lam and Moy (1996a), Kwak et al. (2002), Lam and Moy (2003) conclude that the LP approach for two-group classification problems performs as good as the statistical classification approaches and in many cases even better. However, previous research suggests that there is no single method that clearly out performs all methods in all problem situations (Kiang 2003). For instance, (Asparoukhova and Krzanowskib 2001) show that for small sample sizes, the MP approaches provide best classifiers compared with statistical approaches such a LDF, which provide effective classifiers for large sample sizes. In general, research on two-group classification problems suggest that under varying data characteristics such as presence of outliers, varying sample sizes, non-linearity, non-normality, homoscedasticity, etc., different methods perform differently and emphasize a need for hybrid classifiers to overcome biases in data (Kiang 2003).

There has been very few research studies aimed at three-group classification in the past decade or so. And those that exist are not very comprehensive to judge the efficacy of the MP approaches under robust experimental conditions. Lam and Moy (1996b) propose a LP model for three-group classification problem, which minimizes the sum of individual deviations of the classification scores from their group mean classification scores. The model divides the classification process into two steps: the first constitutes the determination of attribute weights, and the second simultaneously determines the cut-

off scores for the different classification functions, which the authors claim provide better estimates of the group boundaries. The author's compare their models (hit rate = 68%) with Fisher's linear discriminant function (FLDF) (hit rate = 66.67%) and LP approach by Freed and Glover (1986a) (hit rate = 61.33%) by using three examples of small to moderate size datasets to show that the proposed method have an advantage over other methods. However, the study does not address the impact of large datasets, outliers, and other data and group characteristics on the performance of their model. In their research studies, Pavur and Loucopoulos (1995), Loucopoulos and Pavur (1997), Loucopoulos (2001), the authors propose three-group classification MIP models which 'minimize the sum of deviations'. Pavur and Loucopoulos (1995) compare few variants of MIP models and compare it with sequential pairwise approach. The authors conclude that the proposed MIP models yield lower misclassifications than the sequential pairwise approach. However, the authors conclude their results on the basis of the performance of the models on training set only with very small datasets. Loucopoulos (2001) proposes an MIP model for minimization of misclassification costs in three-group problem. The author concludes that the proposed model (hit rate = 97.65%) with equal misclassification costs performs better than both FLDF (hit rate = 89.41%) and Smith's quadratic discriminant function (QDF) (hit rate = 95.29%). The author uses moderate size datasets and states that the computational times for large group sizes and high group overlap could be intensive and prohibitive, which is a drawback of MIP models. Despite these efforts, the development of classification models using MIP formulations still remain a difficult task for large reference sets. In this study, we use moderate to large datasets with varying

data characteristics. Hence, we do not consider MIP model as a potential method for comparison with other techniques used in our study.

1.4 Summary

As stated earlier, this research aims to fill the gaps that presently exist with regards to multi-group classification problems discussed thus far. The current study will fundamentally evaluate the efficacy of LP approaches to multi-group classification problems and compare it with a hybrid technique, statistical methods such as discriminant analysis (DA) and multinomial logistic regression (LR), a neural network (NN), and k-NN. Additional tests will be conducted using various group and data characteristics to test the performance of the models. The remainder of this dissertation will be structured as follows:

The second chapter presents the literature review for the classifications methods used in this study, namely, DA, LR, NN, k-NN and LP approaches (MP1 and MP2), and describe our proposed methodology, i.e. an integrated method. This chapter also highlights the existing research findings and how our study builds upon their foundation. The third chapter discusses model assumptions, hypotheses, and performance measures. The fourth chapter describes the computational experiments. In the fifth chapter, we present the results of our study, whereas in the sixth chapter we discuss conclusions followed by limitations of the study and future research in chapter seven. The chapter eight discusses the research contributions of this study.

CHAPTER 2: CLASSIFICATION METHODS AND LITERATURE REVIEW

This chapter discusses the six methods used in the study, our proposed methodology, and their literature review. The review provides us the basis for forming our hypotheses regarding the possible link between data characteristics and method performances. The review also provides us a deeper insight into the proposed integrated method. There are many variations for each of the methods, especially, the statistical methods (distance metric, quadratic function (QDA), etc.), neural nets (learning rate, weight decay, etc.) and k-NN algorithm (distance metric, value of k, etc.). For this study, however, we restrict our comparison to the basic versions of these methods to maintain genuine characteristics of the original algorithm (Kiang 2003).

2.1 Discriminant Analysis (The Mahalanobis Distance)

The objective of a discriminant analysis (or DA) is to classify objects, by a set of independent variables, into one of two or more mutually exclusive and exhaustive categories. For example, on the basis of an applicant's age, income, length of time at present home, etc., a credit manager wishes to classify this person as either a good or poor credit risk. For the sake of simplicity we will limit this discussion to two-group classifications, later we will comment on n-group discriminant analysis. For notation, let

X_{ji} = i^{th} individual's value of the j^{th} independent variable

b_j = discriminant coefficient for the j^{th} variable

Z_i = i^{th} individual's discriminant score

$Z_{critical}$ = critical value for the discriminant score

Let each individual's discriminant score Z_i be a linear function of the independent variables. That is,

$$Z_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni} \quad (1)$$

The classification procedure follows:

if $Z_i > Z_{critical}$, classify Individual i as belonging to Group 1,

if $Z_i \leq Z_{critical}$, classify Individual i as belonging to Group 2.

The classification boundary will then be the locus of points, where

$$b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni} = Z_i$$

When n (the number of independent variables) = 2, the classification boundary is a straight line. Every individual on one side of the line is classified as Group 1, on the other side, as Group 2. When n = 3, the classification boundary is a two-dimensional plane in 3-dimensional space, the classification boundary is generally an n-1 dimensional hyperplane in n space (Morrison 1969).

In discriminant analysis, the objective of the Mahalanobis approach is to construct a locus of points that are equidistant from the two group centroids. The distance, which is

adjusted for the covariance among the independent variable, is used to determine a posterior probability that can be used as the basis for assigning the observation to one of the two groups. Thus, although the discriminant function is linear in nature, the procedure also provides a probability of group membership, i.e., a nonlinear function of the independent variables in the model. When this probability of group membership corresponds to the probability of choice, effectively we have a choice model with a different functional form (Lawrence et al. 2007, Lawrence et al. 2008).

2.2 Logistic Regression

Logistic regression (or LR), a statistical modeling method for categorical data has expanded from its origins in biomedical research to fields such as business and finance, engineering, marketing, economics, and health policy (Meyers et al. 2006). The availability of sophisticated statistical software and high speed computing has further increased the utility of logistic regression as an important statistical tool.

Logistic regression is particularly suitable for estimating categorical (dichotomous or polytochomous) dependent variables using maximum likelihood estimation (MLE) procedure. Logistic regression models use MLE as their convergence criterion. Logistic regression allows one to predict dichotomous outcome such as presence / absence, success / failure, buy / don't buy, default / don't default, and survive / die. The independent variables may be categorical, continuous or a combination of the both. We can think of categorical variable as dividing the observations into several classes. For example, if Y denotes a recommendation on holding / selling / buying a stock, then we

have a categorical variable with 3 categories. We can think of each stock in the dataset as belonging to one of the three classes: the "hold" class, the "sell" class and the "buy" class. Logistic regression has found two broad applications in applied research: classification (predicting group membership) and profiling (differentiating between two groups based on certain factors) (Tansey et al. 1996, Shmueli et al. 2006).

In general, the logistic regression model has the form

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n = x\beta \qquad (1)$$

Where p is the probability of outcome of interest, β_0 is an intercept term, β_i is the coefficient associated with the corresponding dependent (explanatory) variable x_i, $x = (1, x_1, x_2, ..., x_n)$ and $\beta = (\beta_0, \beta_1, ..., \beta_n)'$.

The probability of outcome of interest, p is expressed as a non-linear function of the predictors in the form

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \qquad (2)$$

The equation (2) ensures that the right hand side will always lead to values within the interval (0, 1). This is called the *logistic response function*.

In the equation (1), the expression

$\frac{p}{1-p} = odds$, which can be rewritten as $p = \frac{odds}{1 + odds}$ (3)

Hence, in logistic regression, one estimates the log of probability odds also known as the *logit* by a linear combination of the predictor variables. The *logit* takes on values from $-\infty$ to $+\infty$.

Taking exponentials of both sides of equation (1) leads to

$$p = \frac{e^{x\beta}}{1 + e^{x\beta}}. \qquad (4)$$

In our study, we use multinomial logistic regression (or MLR) with cumulative logit method. For example, in our financial services segmentation, we have four ordinal classes: prime, highly valued, price shoppers, and no buyers. We denote them by 0 = no buyers, 1 = prime customers, 2 = highly valued customers, and 3 = price shoppers. We look at the cumulative probabilities of the class membership for segmentation. The probabilities that are estimated by the model are $P(X \le 0)$, i.e. the probability of a customer being price shopper and $P(X \le 1)$, i.e. the probability of a price shopper or highly valued customer. The three non-cumulative probabilities of class membership can be easily derived from the two cumulative probabilities:

$$P(X = 0) = P(X \le 0)$$
$$P(X = 1) = P(X \le 1) - P(X \le 0)$$
$$P(X = 2) = P(X \le 2) - P(X \le 1) - P(X \le 0)$$
$$P(X = 3) = 1 - P(X \le 2)$$

Hence, for three classes and three independent variable case, we would have

$$P(X=0)=P(X\le 0)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}}$$

$$P(X=1)=P(X\le 1)-P(X\le 0)=\frac{1}{1+e^{-(\alpha_0+\alpha_1x_1+\alpha_2x_2+\alpha_3x_3)}}-\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}}$$

$$P(X=2)=P(X\le 2)-P(X\le 1)-P(X\le 0)$$
$$=\frac{1}{1+e^{-(\gamma_0+\gamma_1x_1+\gamma_2x_2+\gamma_3x_3)}}-\frac{1}{1+e^{-(\alpha_0+\alpha_1x_1+\alpha_2x_2+\alpha_3x_3)}}-\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}}$$

$$P(X=3)=1-P(X\le 2)=\frac{1}{1+e^{-(\gamma_0+\gamma_1x_1+\gamma_2x_2+\gamma_3x_3)}}$$

The above formulation assumes that the explanatory variables are multivariate normally distributed with equal covariance matrices, explanatory variables are independent and dichotomous zero-or-one variables, or some are multivariate normal and some dichotomous. This makes logistic regression relatively robust compared to linear regression. Also logistic regression models do not assume homoscedasticity between the dependent and independent variables. Under non-normality of explanatory variables, which is very often the case with real-life data, the MLE compares slightly better than, say, linear discriminant estimators. This has been discussed extensively in the literature and is summarized in the excellent text book by Hosmer and Lemeshow (2001).

Logistic regression has been extensively used in family studies and social sciences (Morgan and Teachman 1988). However, we restrict our literature review for logistic regression analysis to business and related fields. One of the earliest applications of logistic regression in business studied models of consumer credit behavior for credit-granting decisions based on a scoring experiment (Wiginton 1980). The paper compares MLE logit model with linear discriminant model and concludes that logit model yields

parameter estimates, which give higher proportion of correct classifications in the scoring experiment. Ball and Tschoegl (1982) model the decision to establish a branch or a subsidiary as a binary choice using the data on the foreign direct investment (FDI) behavior of foreign banks in Japan and California. The results are consistent for both linear discriminant analysis and logistic regression. Walking (1985) develop a logistic regression model for the prediction of tender offer outcomes based on the compensation offered to shareholders. The model correctly classifies 79.6 percent of the 108 offers of estimation sample; however the predictive accuracy of the model on the validation sample of 50 offers is somewhat lower at 60 percent.

Allenby and Lenk (1994) use logistic regression to understand the purchase behavior of households. Kumar et al. (1995) compare a neural network and logistic regression using data collected on the decisions by supermarket buyers whether to add a new product to their shelves or not. The results suggest that logistic regression compares favorably with a neural network in some cases and has superior solution technique and better interpretability. Dasgupta et al. (1994) compare the performances of a neural network and logistic regression with respect to their ability to identify customer segments for an investment product. The results indicate no significant difference between the two models performances contrary to the findings in financial industry applications about the superiority of a neural network. Gan et al. (2005) compare a neural network with logistic model on consumer's banking choices between electronic banking and non-electronic banking. The results indicate that logistic model is accurate in consumer's choice

prediction with overall above 90 percent correct. However, the logistic model produces higher Type I error compared with a neural network model.

In almost all of above studies, logistic regression method is used for binary classification and mostly compared with a neural network technique. In this study, we use the logistic regression model for four group classification problems.

2.3 Neural Networks

Neural networks, also called artificial neural network (ANNs), are models for classification and pattern recognition capabilities. ANNs were designed to model the functioning of human brain, where neurons are inter-connected and learn from experience. We use ANNs for our research for two reasons. First, the ability of the neural networks to decipher and solve nonlinear relationships problems. Second, research over last two decades indicate a neural network may achieve better classification and prediction compared to standard statistical methods (Sharda 1994). This has been corroborated by a number of successful ANN applications such as bankruptcy prediction (Odom and Sharda 1990), bank failure prediction (Tam and Kiang 1990), and market segmentation (Fish et al. 1995, Zahavi and Levin 1997a, Hruschka and Natter 1999) to name a few.

Neural networks structure captures complex relationships between the predictor variables and the response variable through a layer of neurons. Some have one layer – single-layer neural networks (SLNN) and some have more – multilayer neural networks

(MLNN). While various neural networks architectures have been reviewed in the literature, the most successful applications in classification and prediction have been *multilayer feedforward networks*. The layer where input patterns are applied is the *input layer.* The layer from which an output response is desired is the *output layer.* In the case of a binary outcome, the network has only one output node. Layers between the input and output layers are known as *hidden or transfer layers*, because their outputs are not readily observable.

Figure 1 shows a simple multilayer feedforward network comprising of nodes and arrows. The nodes in the network represent neurons while the arrows indicate communication path associated with a synaptic strength or weight value. The arrows connect neurons from one layer to next layer and do not leapfrog layers. The outputs of node in a layer are inputs to the nodes in next layer.

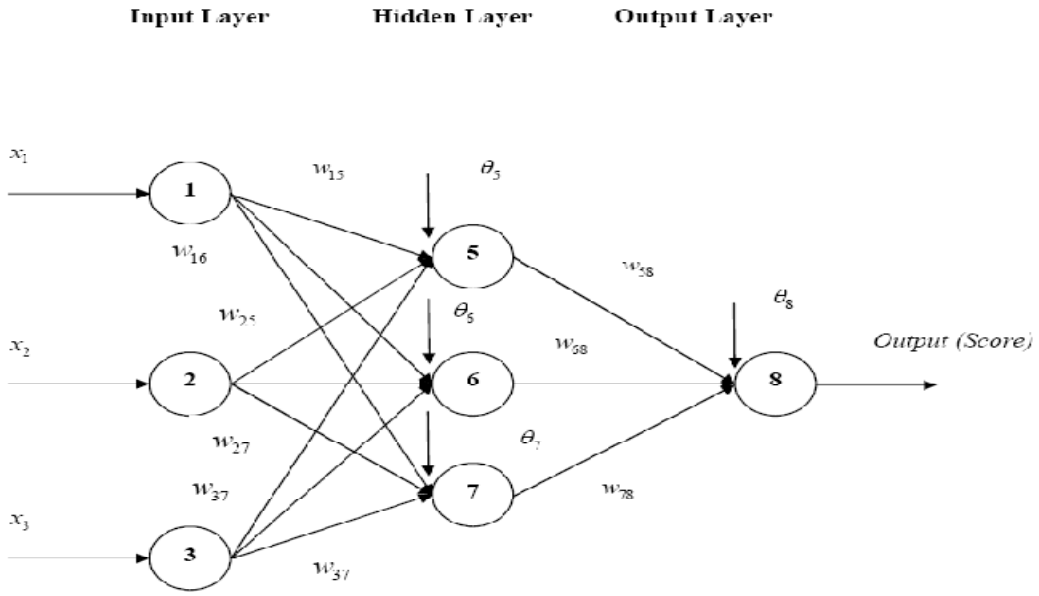

Figure 1: A Diagram of Neural Network

The values on the connecting arrows are called weights, and the weights on the arrow from node i to node j are denoted by $w_{i,j}$. The bias terms, denoted by θ_j, serve as an intercept for the output from node j. The predictor values are supplied as inputs to the input nodes. Their output is the same as the input. If there are p predictors, we have p nodes in the input layers. The outputs of the input nodes forms input to each of the nodes in the hidden layer. There may be more than one hidden layer depending upon the user's objectives. In such case, the output of one hidden layer becomes input to the next hidden layer. However, White (1989) has shown that almost any function can be approximated by a neural network with single hidden layer if the number of hidden nodes is sufficient. Thus, the number of hidden nodes is more critical than the number of hidden layers in a network. To compute the output of the hidden layer, we calculate a weighted sum of the inputs, and then apply certain transfer function to it (Kartalopoulos 1996).

Consider Figure.1 again. Suppose there are p inputs ($p+1$ including the bias term) $x_1, x_2, ..., x_p$. We compute the output of node j by taking the weighted sum $\theta_j + \sum_{i=1}^{p} w_{ij} x_i$, where $\theta_j, w_{1,j}, ..., w_{p,j}$ weights that are set randomly and then adjusted as the network "learns". We take a function f, called a transfer function of this output. The transfer function can be a monotone function such as a linear function ($f(x) = ax$), an exponential function ($f(x) = e^{ax}$), a function producing output between -1 and 1 ($f(x) = \tanh(x)$), and a logistical / sigmoidal function ($f(x) = \frac{1}{1+e^{-x}}$).

The logistic function is commonly used as transfer function because it is a differentiable monotonic function which asymptotically approaches some limit in either direction and thus supplies a good approximation to the threshold function (Hertz et al. 1991). As in logistic regression, the output value is between zero and one.

Using a logistic function we can now write the output of node j in the hidden layer as:

$$output_j = f(\theta_j + \sum_{i=1}^{p} w_{ij} x_i) = \frac{1}{1 + e^{-(\theta_j + \sum_{i=1}^{p} w_{ij} x_i)}}.$$

The learning occurs through the adjustment of path weights and the intercept or bias terms, the values of which are typically initialized to small numbers in the range 0.00 ± 0.05. A neural network model is first trained on a set of input-output using training set data. Training the model means estimating the weights that leads to the best predictive results. The most common method used for the adjustment is an algorithm called the *back propagation*. In this method, the weights are adjusted to minimize the squared difference between the model output and the desired output. The adjustments are based on gradient descent algorithm. Among the many excellent books on neural networks are (Shmueli et al. 2006, Hertz et al. 1991, Fausett 1994).

Despite several advantages such as good predictive performance, high tolerance to noisy data, and its ability to capture and solve complex relationships problems, neural networks have some disadvantages. First, a neural network does not provide insight into the structure of relationship between the predictor and response variables. Second, a neural network does not have built-in variable selection mechanism and hence there is a need to

evaluate the importance of adding predictor variables to the model using other statistical methods. Third, a neural network relies heavily in having sufficient data for training purposes, otherwise the model performs poorly. Fourth, a choice of improper hidden layers or nodes may cause overfitting, that is, the model could get caught in a local rather than a global minimum. Finally, a neural network takes relative higher computational time, which increases with the number of predictors (Uysal and Roubi 1999, Shmueli et al. 2006).

Literature concerning artificial neural networks is replete with its application to business and finance. Furthermore, most classification applications of a neural network pertains (Pendharkar 2002, Kaefer et al. 2005) to two group classification, whereas our examples focus on multiple group classification. One of the first applications of a neural network to classification was by Dutta and Shekhar (1988). They successfully applied neural network to classify industrial bonds based on their risk ratings. Thereafter, Odom and Sharda (1990) developed a neural network model for predicting corporate bankruptcy. They also used a multivariate discriminant analysis technique to compare the performance of neural networks. Their results indicate that neural network outperformed discriminant analysis for both training and validation data sets. Tam and Kiang (1990), Tam and Kiang (1992) apply a neural network to predict bank failures and empirically demonstrate the efficacy of neural networks models to other classification methods. Sharda (1994) surveyed application of neural networks in OR, optimization and statistical methods. The survey results for statistical methods focus on applications of a neural network to firm failure prediction, bank failure prediction, bond rating, and fraud

prevention including others. Lacher et al. (1995) use a neural network to classify financial health of a firm. Fish et al. (1995) apply a neural network to industrial market segmentation (both two-group and three-group classification) using seven input variables or predictor and two hidden layers. The methodology achieved higher hit rates for both training and validation samples. West et al. (1997) use a neural network to predict the outcome of noncompensatory choice rule and to predict consumer perceptions and patronage behavior toward three nationwide mass-merchandise retailers. There are myriad classification applications of a neural network in the literature. Brockett et al. (1997) use a neural network to develop early warning system to predict insolvency for property and casualty insurers. Neural networks application to target marketing is studied by Zahavi and Levin (1997a), Zahavi and Levin (1997b). Zhang et al. (1999) apply a neural network for bankruptcy prediction by using cross-validation method. Some of the other applications of a neural network found in the literature are tourism demand analysis (Uysal and Roubi 1999), market segmentation (Natter 1999, Kim et al. 2003, Bloom 2004, Bloom 2005), bankruptcy prediction using modular neural networks (Nasir et al. 2000), and classify credit risks (Hand 2001).

This study intends to add to the neural network literature in that this would be among the few studies conducted so far on multi-group classifications. One of the unstated objectives of this study is to examine whether or not a neural network can provide some practical benefits with respect to multi-group classification problems. A series of simulations are used to attack this objective, using data generated with various distributions and correlations.

2.4 kth-Nearest neighbor (k-NN)

The k-nearest neighbor decision rule (k-NN) is a commonly used classification algorithm in statistical pattern recognition (Devijver and Kittler 1982, pp 69-127). The idea in k-Nearest Neighbor methods is to identify k observations in the training dataset that are similar to a new record that we wish to classify. We then use these similar (neighboring) records to classify the new record into a class, assigning the new record to the predominant class among these neighbors. Denote by $(x_1, x_2, ..., x_p)$ the values of the predictors for this new record. We look for records in our training data that are similar or "near" to the record to be classified in the predictor space, i.e., records that have values close to $x_1, x_2, ..., x_p$. Then, based on the classes to which those proximate records belong, we assign a class to the record that we want to classify (Shmueli et al. 2006).

Unlike, the classical statistical methods, which make assumptions about the relationship between the response (Y) and predictor variables ($x_1, x_2, ..., x_p$), k-NN is a non-parametric classification method that make no such assumptions. It relaxes the normality assumption and does not require a functional form as required in DA and logistic regression. Instead, this method draws information from similarities between the predictor values of the records in the data set.

The central issue in k-NN is how to measure "distances" or "closeness" between records on their predictor values. Generally, the Euclidean distance and the city-block distance metrics depicted below are employed to calculate distances between two p-dimensional records: $X' = [x_1, x_2, ..., x_p]$ and $U' = [u_1, u_2, ..., u_p]$:

(1) Euclidean distance

$$d(x,u)=\sqrt{(x_1-u_1)^2+(x_2-u_2)^2+\ldots+(x_p-u_p)^2}=\sqrt{(x-u)'(x-u)}$$

(2) City-block distance

$$d(x,u)=\sum_{i=1}^{p}\left|x_i-u_i\right|$$

A number of investigators have considered the question of how best to measure distance: approaches have included global metrics (Fukunaga and Flick, 1984), local metrics (Short and Fukunaga 1980), metrics that are specific to the problem (Simard et al. 1993) and so on. By far the most common metric, though, has been Euclidean distance.

After computing the distances between the record to be classified and existing records, an appropriate k needs to be chosen for the specific problem, which is a tough task because too large (or too small) k may result in non generalizing classifiers. Generally speaking, if k is too low, we may be fitting to the noise in the data. However, if k is too high, we will miss out on the method's ability to capture the local structure in the data, one of its main advantages. In other words, we want to balance between overfitting to the predictor information and ignoring this information completely. A balanced choice depends greatly on the nature of the data. The more complex and irregular the structure of the data, the lower the optimum value of k. Typically, values of k fall in the range between 1 and 20. Often an odd number is chosen, to avoid ties. We choose that k which has the best classification performance, which is determined by using the training data to classify the

records in the validation data, and then compute error rates for various choices of k (Shmueli et al. 2006).

In addition to being a non-parametric method, k-NN has other advantages, such as its ability to perform well in presence of a large enough training set and the minimal time required to find the parameters from the training data. However, k-NN has two main disadvantages: first, the time required to find the nearest neighbors in a large data set can be prohibitive, and second, the number of records required in the training set to qualify as large increases exponentially with the number of predictors p. This is because the expected distance to the nearest neighbor goes up dramatically with p unless the size of the training set increases exponentially with p.

K nearest neighbor classifier (K-NN) is widely discussed and applied in pattern recognition and machine learning due to several interesting features, such as good generalization and easy implementation. Although simple, it is usually able to match, and even beat, more sophisticated and complex methods. Duda and Hart (1973) showed that k-NN can be used to obtain good estimates of the Bayes error and its probability of error asymptotically approaches the Bayes error. Wettschereck and Dieterich (1995) compared nested generalized exemplar (NGE) with K-NN algorithms, and found that k-NN is a fairly robust and effective classifier compared with the nearest hyperrectangle algorithm, an inductive method based on the (NGE) theory (Salzberg 1991).

2.5 Linear Programming (Mean minimization)

The application of LP methods to discriminant problem gained momentum after Freed and Glover (1981a) proposed simple linear programming method. The method identified a weighting scheme to establish a critical value or cutoff point that served as a breakpoint between two groups – successful and unsuccessful. Thereafter, Freed and Glover (1981b) proposed a set of interrelated goal programming formulations. They proved the potential of these formulations with the help of a simple example of assigning credit applicants to risk classifications. Freed and Glover (1987) study two-group problem involving both normal and non-normal populations. The authors propose three contrasting LP formulations: MMD (minimize maximum deviation), MSID (minimize the sum of interior distances), and MSD (minimize sum of deviations) and compare those with classical (Fisher) discriminant procedure. The results indicate that MSD is most effective in correctly classifying population group members for both normal and nonnormal cases. For a classification problem with q criteria, with X, a ($n \times q$) matrix representing the criterion scores of a known sample of n objects from two groups, G_1 and G_2, the Minimize the Sums of Deviations model (MSD) formulation is as below:

$$\text{Min: } \sum_{i=1}^{n} d_i \qquad (1)$$

s.t.

$$\sum_{j=1}^{q} w_j x_{ij} + d_i \geq c, \qquad \text{for } i \in G_1 \qquad (2)$$

$$\sum_{j=1}^{q} w_j x_{ij} - d_i \leq c, \qquad \text{for } i \in G_2 \qquad (3)$$

where,

x_{ij} = value of the j^{th} criterion for the i^{th} object in the sample,

w_j = attribute weights, for $j = 1, ..., q$

c = cut-off score

w_j and c are unrestricted in sign, and

$d_i \geq 0$ for all i, is the deviation of individual objects from cut-off scores

Freed and Glover (1982), Glorfeld and Gaither (1982), Glover et al. (1988), Lee and Ord (1990), Glover (1990), Joachimsthaler and Stam (1990), Ragsdale and Stam (1991) discuss modifications such as additional variables and/or normalization constraints to linear programming approaches and conclude the superiority of such approach over classical statistical techniques. Bajgier and Hill (1982) demonstrate the efficacy of LP approaches using mixed integer, linear goal programming formulations. Markowski and Markowski (1987) extend the above to include qualitative variables. The results indicate improvement in both the LP approach and Fisher's discriminant method with the latter being more preferable. Mahmood and Lawrence (1987) compare nonparametric and parametric discriminant analysis techniques through an empirical study of financial data of companies. The paper classified companies into two groups – bankrupt and non-bankrupt companies. The results indicate that nonparametric approaches such as rank discriminant, log-linear with the exception of linear programming perform better than the parametric approaches in classifying the data into correct groups.

In a departure from the previous research, Joachimsthaler and Stam (1988) compared four discriminant models: Fisher linear discriminant function (FLDF), Smith's quadratic

discriminant function (QDF), the logistic discriminant model and a linear programming (LP) model based on the expected error rates. None of the methods produced significantly lowers rates of misclassification under nonnormality barring QDF, under certain conditions. Rubin (1990) achieve somewhat similar results while comparing fifteen linear programming models under normally distributed datasets. Lam et al. (1996a) propose a two stage LP approach that obtains a set of attribute weights in first stage and determines cutoff score for classification purposes in the second stage. The authors claim this approach obtains more stable classification functions across different samples compared to existing methods. Lam et al. (2003) extend the above to solve the multi-group classification problem. The authors introduce weighted linear programming model (WTLP) and compares it with other discriminant approaches such as MSD, cluster based linear programming (CBLP), and Fisher linear discriminant function (FLDF). Through a simulation experiment the authors claim that the classification performance of WTLP is superior to other methods. Most of the above research delves on two-group classification problems.

Gehrlein (1986) proposes a formulation for the multi-group case. It unfortunately requires a multitude of binary variables in order to identify the optimal division of segments of the decision space among the various groups, rendering its implementation infeasible in practice for many real-size data sets.

Freed and Glover (1981b) state that minimizing the sum of deviations (MSD) formulation, which is one of the most widely used linear programming (LP) formulations

for solving the classification problem, can easily be generalized to the multi-group classification problem by sequentially solving for the optimal separating hyperplanes between the pairs of groups. One problem with this approach, however, is that the resulting classification rules may not cover each segment of the decision space. Moreover, the pairwise estimation of hyperplanes leaves much to be desired, because it may lead to suboptimal overall classification results. Another approach they suggested is to convert the classification problem to $m(m-1)/2$ distinct two-group problems (where m = number of designated groups), where each problem is solved separately.

Pavur and Loucopoulos (1995), Lam and Moy (1996b), Gochet et al. (1997) also study the three-group classification problems using small to moderate data sets. Pavur and Loucopoulos (1995) study the three-group classification problem using MIP approach. They use small data sets for their experiments and conclude that their models perform better than some of the statistical approaches in achieving lower misclassification rates.

Lam and Moy (1996b) extend two-group classification model developed by (Lam et al. 1996a) to solve multi-group classification problem.

Let the mean of the j^{th} variable for $k = 1, 2, ..., m$ be $\bar{x}_j(k) = \dfrac{\sum_{i \in G_k} x_{ij}}{n_k}$

where n_k = number of observations in G_k, and m = number of designated groups.

Let n be the total number of observations, $n = n_1 + n_2 + ... + n_m$

We consider a classification problem with q variables and n, the total number of observations in the sample. For each pair of (u,v), where $u = 1,...,m-1, v = u+1,...,m$, the Minimize the Sum of Deviations model (MSD) formulation is:

$$\text{Minimize} \sum_{i \in G_u, i \in G_v}^{n} d_i \qquad (1)$$

s.t.

$$\sum_{j=1}^{q} w_j (x_{ij} - \bar{x}_j(u)) + d_i \geq 0, \qquad \forall i \in G_u \qquad (2)$$

$$\sum_{j=1}^{q} w_j (x_{ij} - \bar{x}_j(v)) - d_i \leq 0, \qquad \forall i \in G_v \qquad (3)$$

$$\sum_{j=1}^{q} w_j (\bar{x}_j(u) - \bar{x}_j(k)) \geq 1, \qquad (4)$$

where,

w_j = weights, for $j = 1,...,q$ are unrestricted in sign

x_{ij} = value of j^{th} variable for the i^{th} observation in the sample

$d_i \geq 0$ for $i \in G_u$ and $i \in G_v$, is the deviation of an individual observation from the cut-off score.

The objective function (1) minimizes the sum of all the deviations (MSD). The constraints (2) and (3) force the classification scores of the objects in G_k to be as close to the mean classification score of group k (k =1, 2,..., m) as possible by minimizing d_i

where $i \in G_k$. The constraint (4) is a normalization constraint to avoid trivial values for discriminant weights.

For each pair (u,v), we use the w_j values obtained from the LP solution of pair (u,v) to compute the values of the classification scores, S_i of the observations in G_u and G_v. Then all the cut-off values, C_{uv}, where $u = 1,...,m-1, v = u+1,...,m$, are determined by solving the following LP problem,

$$\textbf{Minimize} \sum_{i \in G_u} \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} d_{iuv} + \sum_{i \in G_v} \sum_{v=1}^{m-1} \sum_{v=u+1}^{m} d_{iuv} \tag{5}$$

s.t.

$$S_{iuv} + d_{iuv} \geq c_{uv}, \qquad \textbf{for } u = 1,...,m-1, v = u+1,...,m, i \in G_u \tag{6}$$

$$S_{iuv} - d_{iuv} \leq c_{uv}, \qquad \textbf{for } u = 1,...,m-1, v = u+1,...,m, i \in G_v \tag{7}$$

where all c_{uv} are unrestricted in sign and all $d_{iuv} \geq 0$, and $S_i = \sum_{j=1}^{q} w_j x_{ij}$, for $i \in G_k$

The author's compare their models with FLDF and MP approach of Freed and Glover (1986a) by using three examples of small to moderate size datasets to show that the proposed method have an advantage over other methods.

Gochet et al. (1997) propose a nonparametric linear programming formulation for the general multi-group classification problem. The authors, with the help of several small

examples, show that their proposed multi-group LP approach offers a robust alternative to both Fisher's parametric method and non-parametric *k*-nearest neighbor method.

Clearly, in all of the research on multi-group classifications to date, the robustness of the proposed classification methods, with respect to various data conditions, must be ascertained. This related to the mixed success for the two-group case.

2.6 Linear Programming (Median Minimization)

In this study, we extend (Bal et al. 2006) for two-group classification to multi-group classification problems. This extension is based on (Lam and Moy 1996b), which, proposes a model regarding minimization of deviations from the group means. However, for non-normal distribution, this model loses efficiency in respect of hit ratio. For the samples draw from the non-normal or skewed distributions, the median is a much more suitable descriptive statistic than the mean. For two-group classification problems (Bal et al. 2006) show that the performance of their model is better than both some important classification in literature and the model suggested by (Lam and Moy 1996b).

Let n be the total number of observations, $n = n_1 + n_2 + ... + n_m$

We consider a classification problem with q variables and n, the total number of observations in the sample. For each pair of (u, v), where $u = 1, ..., m-1, v = u+1, ..., m$, the Minimize the Sum of Deviations model (MSD) formulation is:

Minimize $\sum_{i \in G_u, i \in G_v}^{n} d_i$ (1)

s.t.

$$\sum_{j=1}^{q} w_j (x_{ij} - med_j(u)) + d_i \geq 0, \quad \forall i \in G_u \qquad (2)$$

$$\sum_{j=1}^{q} w_j (x_{ij} - med_j(v)) - d_i \leq 0, \quad \forall i \in G_v \qquad (3)$$

$$\sum_{j=1}^{q} w_j (med_j(u) - med_j(v)) \geq 1, \qquad (4)$$

where,

w_j = weights, for $j = 1, ..., q$ are unrestricted in sign

x_{ij} = value of j^{th} variable for the i^{th} observation in the sample

$med_j(u)$ = the median of the j^{th} variable in group *u*.

$med_j(v)$ = the median of the j^{th} variable in group *v*.

Pair (u, v) = any group u =1, 2... m-1 and v = u+1, ..., m (what this means is that we solve an LP for each pair of group to generate weights)

$d_i \geq 0$ for $i \in G_u$ and $i \in G_v$, is the deviation of an individual observation from the cut-off score.

The objective function (1) minimizes the sum of all the deviations (MSD). The constraints (2) and (3) force the classification scores of the objects in G_k to be as close to the mean classification score of group k (k =1, 2,..., m) as possible by minimizing d_i

where $i \in G_k$. The constraint (4) is a normalization constraint to avoid trivial values for discriminant weights.

For each pair (u,v), we use the w_j values obtained from the LP solution of pair (u,v) to compute the values of the classification scores, S_i of the observations in G_u and G_v. Then all the cut-off values, C_{uv}, where $u = 1,...,m-1, v = u+1,...,m$, are determined by solving the following LP problem,

$$\text{Minimize } \sum_{i \in G_u} \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} h_{iuv} + \sum_{i \in G_v} \sum_{v=1}^{m-1} \sum_{v=u+1}^{m} h_{iuv} \tag{5}$$

s.t.

$$S_{iuv} + h_{iuv} \geq c_{uv}, \qquad \textbf{for } u = 1,...,m-1, v = u+1,...,m, i \in G_u \tag{6}$$

$$S_{iuv} - h_{iuv} \leq c_{uv}, \qquad \textbf{for } u = 1,...,m-1, v = u+1,...,m, i \in G_v \tag{7}$$

where all c_{uv} are unrestricted in sign and all $h_{iuv} \geq 0$, and $S_i = \sum_{j=1}^{q} w_j x_{ij}$, for $i \in G_k$

2.7 Integrated (Hybrid) Method

Previous research on classification indicates that no single method clearly outperforms all methods in all problems (Ostermark 1998). That is, different kinds of methods have their own advantages and defects. So, a method can perform best for one specific problem, but given another problem, another method can work better. This situation is called *selective superiority* (Michie et al. 1994). Therefore, one recommendation is to build classification systems that employ a number of different classification algorithms to improve the classification and prediction accuracy. The systems could be designed to select the right

method or to properly combine different methods to form an integrated or hybrid classifier in response to the different biases in data (Kiang 2003). Hybrid models combine different methods to improve classification and prediction accuracy. The term combined or integrated model is usually used to refer to a concept similar to a hybrid model. Combined models also have been called *Ensembles*. Ensemble improves classification and prediction performance by the combined use of two effects: reduction of errors due to bias and variance (Haykin 1999).

Hybrid models and combined models, terms often used interchangeably, have been developed to improve the classification and prediction accuracy by using several supervised learning methods together. Some studies on hybrid models utilize different supervised learning methods *sequentially* (Hur and Kim 2008). Utgoff (1989) presented a hybrid representation, called a 'perceptron tree', and an associated learning algorithm called the 'perceptron tree error correction procedure' by using the favorable characteristics of a decision tree and linear threshold units (LTUs). The rationale is that the two algorithms complement each other in certain ways, and by properly integrating them into one method, one can draw on the particular strengths of each individual algorithm. Coenen et al. (2000) propose a hybrid model to improve the response rate of direct mailing. They use the C5.0 method for initial classification of buyers and non-buyers and then use case-based reasoning for ranking the classified cases. Carvalho and Alex (2004) suggest a hybrid model that generates decision rules using C5.0 and selects final decision rules with a genetic algorithm. Li and Wang (2004) present a method that

can improve the effectiveness of final classification rules using artificial neural networks and the rough set theory presented by Pawlak (1991).

The above hybrid models use different models with a phased approach. That is, one method is used first in some data mining phase, and the other method is used in a next phase. Another hybrid approach is embedded. That is, a method or technique is embedded into part of a main method and carries out a subtask to improve the performance of the main method (Hur and Kim, 2008). For example, Chen (2003) suggests a hybrid framework for textual classification in text mining using fuzzy theory embedded in a SOM (self-organized map).

In addition to hybrid methods that have tried to combine two completely different methods, hybrid models that use one method in multiple ways have also been studied. Hansen and Salaman (1990) show that the generalization ability of a neural network system can be significantly improved through ensembling a number of neural networks. Indurkhya and Weiss (1998) show the improvement of predicted gain values of the final nodes in decision trees by multiple re-sampling of decision tree induction methods and combination of them using the voting method.

There are also studies on the predictability or classification performance of hybrid or combined models compared with a single method. Kuncheva et al. (1998) presented cases in which prediction accuracy was improved using hybrid models. With combinations of RFM, neural networks, and logistic regression models, Suh et al. (1999)

showed that performance of hybrid techniques improves when the correlation between hybrid models is low. Zhang and Zhang (2004, Chapter 8) explain that a single data mining technique has not been proved appropriate for every domain and data set. Instead, several techniques may need to be integrated into hybrid systems that can be used cooperatively during a particular data mining operation.

An alternative approach is to combine the outputs of different classification methods. Wolpert (1992) introduced stacked generalization, a way to combine the outputs from multiple generalizers trained with multiple partitionings of the original learning set. However, there are no systematic rules that can be used to generate an accurate combination. Breiman (1996a) followed Wolpert's idea of combining predicators instead of selecting the single best method, and proposed stacked regressions method. Stacked regression is a method for forming a linear combination of different predicators to give improved prediction accuracy. In general, improvement occurs when stacking together more dissimilar predictors. Bagging predictors, proposed by Breiman (1996b), is a method for generating multiple versions of a predictor, then obtaining an aggregated predictor by either taking the average over the versions (for numerical output) or using a plurality vote (for classification tasks). Brieman (1996b) reported that prediction accuracy can be improved from 57% to 94% by applying Bagging to the C&RT algorithm and demonstrated that the stability of a procedure has great impact on the improvement achieved through bagging. The author studied the instability of different predictors and concluded that neural networks, classification trees, and subset selection in

linear regression were unstable, while the kth-nearest-neighbor method was stable (Kiang 2003).

This study constructs an integrated (hybrid) classifier by combining two methods in common use – *k*-nearest-neighbor (*k-NN*) and a linear programming (LP) approach. In this scheme we divide the initial feature space up by *k-NN*, and then classify the training set using LP approach. The *k-NN* method acts as a data preprocessing stage, where, it is used to discard the unwanted data i.e., the group Y = 0 (no buyers customer segment, in this study) for final classification using LP approach. This preprocessing step helps us in two ways: first, a major problem of using the *k-NN* is the computational complexity caused by large number of distance computations (Devijver and Kittler 1982). The preprocessing stage helps in reducing this complexity of the initial problem for k-NN by having to classify only two groups i.e., Y= 0 and Y = 1; second, to reduce considerably, the number of constraints required in LP approach. This way, we try to alleviate the disadvantages of both these methods, and at the same time utilize their strengths for improving classification accuracy. For the LP approach part of this hybrid approach, this study utilizes multi-group classification LP method developed by (Lam and Moy 1996b).

Previous researches have studied the problem of partitioning the feature space into different subsets for discrimination of different pattern classes using composite classifiers. The composite classifiers have been found to lead to improved performance in multiclass environments (Kanal and Chandrasekaran 1972, Dasarathy 1973). Dasarathy and Sheela (1979) study the linear/NN classifier composite which ensures that

the computational effort is less than that under NN classifier, and, at the same time, the recognition rate is equal to or better than under each of the components, thereby meeting the objective of improved recognition system performance. Buttrey and Karo (2002) have constructed a hybrid (composite) classifier by combining two classifiers in common use: classification trees and k-NN. The authors divide the feature space up by a classification tree, and then classify the test set items suing the k-NN rule just among those training items in the same leaf as the test item. The authors claim that this reduces the computational load associated with k-NN, and it produces a classification rule that performs better than either trees or the usual k-NN in a number of well known data sets.

CHAPTER 3: MODEL ASSUMPTIONS AND PERFORMANCE MEASURES

One of the focuses of this research is to examine the group and data characteristics that may affect the performance of different classification methods. Since real-world data are usually contaminated (Glorfeld and Kattan 1989, Hample et al. 1986, Stam and Ragsdale 1992), this simulation experiment generated data with various characteristics. The characteristics were selected based on previous research in this area and on the identified strengths and weaknesses of each method. The following provides a detailed description for each data characteristic.

3.1 Data characteristics

3.1.1 Multivariate normal (Symmetric)

One of the drawbacks of parametric methods is the normality assumption of the independent variables. However, real-life datasets seldom follow normal distribution (Eisenbeis 1977). Violations of normality assumptions in parametric methods may lead to a biased and overly optimistic classification rates in the population, and thus limit the usefulness of the model (Kiang 2003). The Kolmogorov-Smirnov test statistics are applied to each of the independent variables in the data set to test for normality (Lilliefors 1967, Dyer 1974).

3.1.2 Non-normal data (Asymmetric)

Since, in practice, data are rarely multivariate normally distributed, we also wish to test the performance of the selected classification procedures when allowing departures from normality. For the non-normal data used in our study, we generate lognormal variables (Ostermark 1998). The choice of lognormal distribution is based on the knowledge that this type of distribution is different from the normal curve in overall shape as well as skewness and kurtosis. Other scenarios for departure from non-normality are possible (Hosseini and Armacost 1994).

3.1.3 Dynamic versus static nature of the problem

Most of the methods examined assume that the population distribution will not change with time. Thus, the models based on historical data are not time-dependent and may be violated at times. Time series analysis is one approach to this type of problem. A time series model tries to account for as much as possible of the regular movement (wavelike functions, trend, etc) in the time series, leaving out only the random error. The method can be applied when there is a time series variable in the problem to be modeled. However, a more complex dynamic system could affect the distributional characteristics of the model over time (Kiang 2003).

3.1.4 Outliers (With / without)

To emulate real life datasets, we introduce outliers in our experiments. The outlier datasets contain 5 per cent observations as outliers generated using the Cauchy

distribution (Ostermark 1998). The use of the Cauchy distribution to generate outliers has been supported in the literature (Hoaglin 1985). Other simpler approaches, such as the generation of observations that are several standard deviations from the mean values of the variables are also possible (Bajgier and Hill 1982).

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

3.1.5 Multicollinearity

Past research suggests that a high degree of correlation among independent variables (multicollinearity) will have adverse effects on the parameter estimates of parametric methods (Meyers et al. 2006). Two methods commonly used to test collinearity are correlation matrix and variance inflation factor (VIF) (Neter et al. 1990). This study tests the models for two levels of correlation – strong, and weak.

Correlation matrices are:

$$\Sigma_{Strong} = \begin{bmatrix} 1.00 & 0.75 & 0.75 \\ 0.75 & 1.00 & 0.75 \\ 0.75 & 0.75 & 1.00 \end{bmatrix} \qquad \Sigma_{Weak} = \begin{bmatrix} 1.00 & 0.25 & 0.25 \\ 0.25 & 1.00 & 0.25 \\ 0.25 & 0.25 & 1.00 \end{bmatrix}$$

3.1.6 Homoscedasticity

The linear discriminant analysis (LDA) requires the covariance equality of multi-groups. We test for the equality of variances by conducting Cochran's test (Neter et al. 1990).

This study tests the models by introducing unequal covariance matrices between four groups to test the impact on performance of various methods. We alter the covariance matrices to reflect different degrees of correlation between discriminating variables (Ostermark 1998).

3.1.7 Sample proportion

Previous research indicates that the sample proportion does affect the prediction accuracy of a discriminant model. For instance, DA models show that when sample proportion differs from the true population, the prediction accuracy becomes very poor (Kiang 2003). However, the predictive accuracy of a logit model is not affected by biased sample proportion due to its non-parametric nature. This study uses the same sample proportion for each group as per the reference data sets. The proportion of the sample in per cent terms is: $n_1 = 40\%, n_2 = 10\%, n_3 = 20\%, n_4 = 30\%$, where, $n = n_1 + n_2 + n_3 + n_4$, the sample size for each replication of training and validation set.

3.1.8 Sample size

Previous research in classification studies suggests that size of training samples not only affects speed of training, but also has an impact on the performance of different classifiers. Sordo and Zeng (2005) show through their empirical study that as sample size increases, both support vector machines and decision trees show a substantial improvement in performance, suggesting a more consistent learning process. For some methods, large sample size is required in order to achieve its maximum prediction

accuracy whereas others may need a relatively small data set. In this study, sample sizes of 100, 200, 400, and 500 are randomly selected from the data set each time for both training and validation purposes.

3.2 Hypothesis testing

In this study, we conduct testing of hypothesis at two levels: the effect of data characteristics on various methods based on the strengths of the methods and the performance of the proposed integrated method vis-à-vis the other methods. The review in Chapter 2 provides us the basis for forming hypotheses regarding the possible link between data characteristics and method performances.

Table 1: Summary of the seven methods with respect to all the data characteristics

Method	Hypotheses
DA	Static scenario. Affected by normality and linearity violations, low correlations, outliers and identical covariances.
Logistic	Static scenario. Affected by sample size, especially when dependent variable has many groups, and low correlations.
Neural nets	Both static and dynamic scenarios. Affected by sample size and outliers.
KNN	Static scenario. Affected by sample size and outliers.
MP1	Static scenario. Affected by linearity violations.
MP2	Static scenario. Affected by linearity violations.
Integrated	Static scenario. Affected by linearity violations and sample size.

3.2.1 Effect of data characteristics

Here, we draw heavily from a previous study by Kiang (2003), which tests the hypotheses regarding the possible link between data characteristics and method

performances. Table 1 summarizes our hypotheses regarding the classification methods based on a review of the classification literature. These results are tested and validated using simulated data sets. The performance difference for each method before and after the change is used to test the hypotheses. Paired *t* tests are used to compare the means of the misclassification rates between the base case and the biases in data (Lam et al. 1996a, Kiang 2003, Bal et al. 2006).

3.2.2 Performance of integrated method versus other methods

Here, we use the paired t-tests to test the difference between the misclassification rates of the integrated methods and the other six approaches: discriminant analysis (DA), logistic regression (LR), neural network (NN), k-NN algorithm, and linear programming approaches (MP1 and MP2), for each of the data characteristic. In all we test 54 hypotheses (9 data characteristics and 6 different methods).

Null hypothesis:

H_0: There is no difference between the mean misclassification rates of the integrated method and the misclassification rates of the method *i,* for each of the nine data characteristic.

$$\mu_{Integrated,j} \leq \mu_{i,j}$$

H_a: The mean misclassification rates of the integrated method are greater than the misclassification rates of the method *i,* for each of the nine data characteristic.

$$\mu_{Integrated,j} > \mu_{i,j}$$

Where,

i = Methods DA, LR, NN, k-NN, MP1 and MP2, and

j = Data characteristics: base case, dynamic, nonlinearity, nonnormal, outliers, strong correlation, unequal covariance, unequal sample proportion, weak correlation.

The details of the simulation experiments are discussed in chapter four and the results in chapter five.

3.3 Performance measures

One of the factors used for performance estimation is the way in which multivariate observations are used to design the classifier and to test its performance. There are four main approaches to use given observations as the design set (i.e. the training set) and as the test set (i.e. the validation set) (Raudys and Jain 1991).

1. The Resubstitution Method R: all observations are used to design the classifier and used again to estimate its performance.

2. The Hold-Out Method H: Suppose the total number of available observations is n*. One portion of the set of observations (the training set containing N observations) is used to design the classifier, and the remaining (n* - N) portion (the validation set) is used to estimate the error rate.

3. The Cross-Validation Method L: In this method, $\binom{n^*}{k}$ classifiers are designed. Each classifier is designed by choosing k of the n* observations as a training set, and its error rate is estimated using the remaining (n* - k) observations. This process is repeated for all distinct choices of k patterns and the average of the error rates is computed. A popular choice for the value of k is k = 1, yielding the well-known leave-one-out method.

4. The Bootstrap Method B: A bootstrap design sample of size N is formed from the N observations by sampling with replacement.

In this study, we use the 'Hold-Out Method (H) for estimating both our performance measures: Misclassification rates and Individual error rates. For all the methods of classification and for each of the data characteristics, we use 60 percent of the sample as training data and the remaining 40 percent as the validation data.

3.3.1 Misclassification rates (Apparent error rates)

One important way of judging the performance for any classification procedure is to calculate its error rates or misclassification probabilities. The performance of a sample classification function can be evaluated by calculating the Actual Error Rate (AER). The AER indicates how the sample classification function will perform in future samples. Just as the optimal error rate, it cannot be calculated because it depends on an unknown density function. However, an estimate of a quantity related to the AER can be calculated.

There is a measure of performance that does not depend on the form of the parent population, which can be calculated for any classification procedure. This measure is called the Apparent Error Rate (APER). It is defined as the fraction of observations in the training sample that are misclassified by the sample classification function.

The APER can be easily calculated from the confusion matrix, which shows actual versus predicted group membership. Previous research in classification and its applications in accounting and finance, and marketing show that confusion matrix is the most frequently used performance evaluation measure (Odom and Sharda 1990, Salchenberger et al. 1992, Tam and Kiang 1992, Altman et al. 1994, Wilson and Sharda 1994, Spear and Leis 1997, Lee et al. 2005, Paliwal and Kumar 2009)

For n_1 observations from Π_1, n_2 observations from Π_2 , n_3 observations from Π_3, and n_4 observations from Π_4, the confusion matrix is given in Table 2 ((Morrison 1969):

Table 2: Calculating Misclassification Rates

Actual Membership	**Predicted Membership**					
		Π_1	Π_2	Π_3	Π_4	
	Π_1	n_{11}	n_{12}	n_{13}	n_{14}	n_1
	Π_2	n_{21}	n_{22}	n_{23}	n_{24}	n_2
	Π_3	n_{31}	n_{32}	n_{33}	n_{34}	n_3
	Π_4	n_{41}	n_{42}	n_{43}	n_{44}	n_4

Where,

n_{11} = number of Π_1 items correctly classified as Π_1 items

n_{12} = number of Π_1 items misclassified as Π_2 items

n_{13} = number of Π_1 items misclassified as Π_3 items

n_{14} = number of Π_1 items misclassified as Π_4 items

n_{21} = number of Π_2 items misclassified as Π_1 items

n_{22} = number of Π_2 items correctly classified as Π_2 items

n_{23} = number of Π_2 items misclassified as Π_3 items

n_{24} = number of Π_2 items misclassified as Π_4 items

n_{31} = number of Π_3 items misclassified as Π_1 items

n_{32} = number of Π_3 items misclassified as Π_2 items

n_{33} = number of Π_3 items correctly classified as Π_3 items

n_{34} = number of Π_3 items misclassified as Π_4 items

n_{41} = number of Π_4 items misclassified as Π_1 items

n_{42} = number of Π_4 items misclassified as Π_2 items

n_{43} = number of Π_4 items correctly classified as Π_3 items

n_{44} = number of Π_4 items correctly classified as Π_4 items

The apparent error rate is thus

APER = $1-\frac{n_{11}+n_{22}+n_{33}+n_{44}}{n}$ or, in other words, the proportion of items in the training set that are misclassified, where, $n = n_1 + n_2 + n_3 + n_4$

The APER is intuitively appealing and easy to calculate. Unfortunately it tends to underestimate the AER, and the problem does not appear unless the sample sizes of n_1, n_2, n_3, and n_4 are very large. This very optimistic estimate occurs because the data used to build the classification are used to evaluate it.

The error rate estimates can be constructed so that they are better than the apparent error rate. They are easy to calculate, and they do not require distributional assumptions. Another evaluation procedure is to split the total sample into a training sample and a validation sample. The training sample is used to construct the classification function and the validation sample is used to evaluate it. The error rate is determined by the proportion misclassified in the validation sample. This method overcomes the bias problem by not using the same data to both build and judge the classification function.

3.3.2 Individual error rates

Previous literature on the classification and prediction analysis suggests there are few studies that delve on the individual error rates. Balancing of error rates is the individual error rates for each group or class of the categorical variable, i.e. we count the number of group 0, 1, 2, and 3 values that are misclassified. Markowski (1990) reported the balancing of error rates for an experimental comparison between a linear programming (LP) approach and Fisher's linear discriminant function (FLDF). The study concludes that FDLF is much more effective, when balance between two types of misclassifications is important.

In this study we wish to examine the error rates for the four groups individually under varying data circumstances. We are particularly interested in a lower misclassification rates for top customer segments, hence, we judge the effectiveness of a method by its ability to classify groups Y = 1, and Y = 2 accurately. For instance, in our example, groups 1 (prime) and 2 (high value) are our top customer segments. We would be interested indentifying the methods that have lower misclassification rates for these two groups. This analysis will make the practitioner aware of the inherent strengths and weaknesses of classification techniques.

CHAPTER 4: COMPUTATIONAL EXPERIMENTS

Our study is restricted to four-group classification with three discrimination variables. We test the robustness of various methods using a financial services segmentation problem with three independent variables and a categorical dependent variable with four customer class. All the independent variables in our example are continuous. The study uses the characteristics of real data sets to simulate (via Monte Carlo simulation) sample runs for experiments.

4.1 Example - Financial services segmentation

This example focuses on segmenting the financial services market for effectively targeting customers who offer higher expected growth in the value of future business. More specifically, this study attempts to develop a discriminant model to classify the customers based on their demographics i.e. age (X_1), income (X_2), and loan activity (X_3) as independent variables. We segment the customers into four ordinal classes: Y = 0, Y = 1, Y = 2, Y = 3, i.e. non-buyers (n_1), prime customers (n_2), highly valued customers (n_3), and price shoppers (n_4), respectively. The prime customers are the ones who have higher income levels and a loan activity commensurate with their income. They form the most desirable targets for the companies offering financial services. The highly valued customer class has income levels and loan activity relatively lower than the prime customers but profitable enough in the long run though with associated risks. The price shoppers are short term customers with lower long term attractiveness but provide

enough volume base. They are also the ones which cost higher to service due to their tendency to base their decisions on short term benefits and price sensitivity. Lastly, the non-buyers are the ones who are not likely to buy the financial services in a short to medium term.

For the financial services example we use an individual's income, loan activity, and age as explanatory variables. The response variable in our model is a multi-group variable which, indicates whether customers are: prime customers, highly valued customers, price shoppers, or non-buyers. To evaluate the performances of all the methods, a Monte Carlo simulation experiment is conducted to generate sample runs, based on the characteristics of a real consumer dataset.

We compare the performances of the discriminant analysis-Mahalanobis (DA), multinomial logistic regression (LR), LP methods based on: minimize the sum of deviations model (MP1) (MSD) (Lam et al. 1996b), and median minimization (MP2) (Bal et al. 2006), k-NN, and the proposed integrated method for the problem of four-group classification.

4.2 Data generation

To test the effect of each data characteristics, a population of 100,000 cases is generated each time. An equal number of cases (25,000) are generated for each category or class, i.e. Y = 0, Y = 1, Y = 2, and Y = 3 groups. The data sets are generated for nine different

data characteristics using Monte Carlo simulation method. To form the training and validation data sets, 125 cases are randomly drawn from each group for a total of 500 cases in each data set. The process is repeated 150 times to form 150 training and validation data sets, respectively, in order to average out the possible bias in any single sample run. The results presented below are the average performances of the 150 runs, both for training and validation.

The following data characteristics describe the biases inserted at each step during the test.

1. Dynamic environment: Again, the same functional form as the base cases is used. Instead of using a constant A_1 as the coefficient of X_1, it is assumed that the coefficient of X_1 changes over time. A sine function is used as part of the coefficient value from 0 to 1 to 0 to -1, then back to 0. Each time, a complete cycle is used to generate 300 examples and then chronologically divided into two sets. The first 150 examples are used for training and the rest are used as validation sample (Kiang 2003).

2. Nonlinearity: A quadratic function is used in this test:

$$Y = A_1 X_1^2 + A_2 X_2^2 + A_3 X_3^2 + \varepsilon,$$

where $X_1 \sim N(\mu_1, V_1)$, $X_2 \sim N(\mu_2, V_2)$, $X_3 \sim N(\mu_3, V_3)$, and $\varepsilon \sim N(0,1)$. Again, $A_1, A_2, A_3, V_1, V_2, V_3, \mu_1, \mu_2$, and μ_3 are constants and were chosen to make four distinct groups.

3. Nonnormal distribution: A data set with lognormal distribution is generated to compare with normally distributed sample (Ostermark 1998). Only positive values are possible for the variable, and the distribution is skewed to the left. Two parameters are needed to specify a log-normal distribution. Traditionally, the mean μ and the standard deviation σ (or the variance σ^2) of log(X) are used. A random variable X is said to have the lognormal distribution with parameters $\mu \in$ □ and σ>0 if ln(X) has the normal distribution. Equivalently, $X = e^Y$ where Y is normally distributed with mean μ and standard deviation σ. The lognormal distribution is used to model continuous random quantities when the distribution is believed to be skewed, such as certain income and lifetime variables.

4. Outliers: The Cauchy distribution is used to generate outliers in the base cases. We insert 5% of the observations as outliers.

5. Strong correlation: To generate data sets with strong correlation between variables X_1, X_2, and X_3, we use strong and weak correlation matrices available in the previous research literature (Lam and Moy 2003).

6. Unequal covariance: Data sets with different covariance matrix for the four groups, i.e. Y = 0, Y = 1, Y = 2, and Y = 3 were generated.

7. Unequal sample proportion: The sample cases are randomly drawn from the same population used in the base case. A sample proportion of 40-30-20-10

percentages are used for groups Y = 0, Y = 1, Y = 2, and Y = 3, respectively, for both training and validation data sets.

8. Sample size: Sample sizes of 100, 200, 400, and 500 are randomly selected from the base case data set each time for both training and validation.

For each data set generated, necessary tests were performed (i.e., plotting scatter plots, normality tests, etc.) to verify the existence of bias in the data. Performance is assessed with respect to the ability of the methods to accurately predict the appropriate class for the validation sample.

Each experiment includes 300 sample runs (150 training runs and 150 validation runs), and the results presented are the average of the 150 runs for training and validation data sets, each. Therefore, there are a total of 300 (sample runs/cell) x 7 (models) x 9 (data characteristics) = 18,900 runs. To test the effect of sample size on model performance, 150 trainings and 150 validation runs were performed for each sample size. Therefore, there are a total of 300 (sample runs/cell) x 7 (models) x 4 (sample sizes) = 8,400 runs, for the sample size effects.

We used Minitab 15.0 to generate all the data sets required for this study. For DA (Mahalanobis), and logistics regression methods, again we use Minitab 15.0. For neural network analysis, and k-NN, we used XLMiner software. For solving both MP1 and

MP2 problems we used Premium Solver software by Frontline Systems. All these software packages are commercially available.

CHAPTER 5: ANALYSIS OF RESULTS AND DISCUSSION

The results of the training and validation data are shown in Table 3 and Table 4, respectively.

Figure **2** plots the misclassification rates of the validation results for the nine data characteristics and groups them by method. Figure 3 shows the classification performance versus the sample size. Due to the complexity of the problem in this study, the possible interaction among factors and the varying degree of biases in each data characteristic was not tested. In order to test all the possible interactions among biases, the experimental design necessary to test these hypotheses will be very complex due to large number of factors involved (Ostermark 1998, Kiang 2003).

For each method, *t*-test is used to test the significance of the performance difference between the base case and each biased sample (see Table 4). Since lower misclassification rates on the validation data are deemed as a good check on the external validity of the classification function, *t*-tests are conducted only on the validation data. The results show that, in general, logistic regression, as indicated by its high APER, performs poorly whereas the integrated method performs relatively better than most other methods, on all the data characteristics for both training and validation data. The results also show that except for degree of correlation and unequal covariance, all other bias factors have either a nonsignificant or an adverse effect on the performance of a method. LR is the most severely affected whereas the LP methods are relatively less affected by the bias factors. The standard deviations of the average number of misclassifications for

the DA, LR, NN show that they are relatively less robust methods compared with k-NN, LP1, LP2, and integrated methods.

Table 3: Misclassification Rates for the Training Data[1]

Method	Base	Static/ Dynamic	Linearity assumption	Normality assumption	Outliers	Strong correlation	Unequal covariance	Sample proportion	Weak correlation
DA	19.92 *(3.06)*	31.61 *(2.52)*	33.58 *(3.03)*	33.06 *(3.18)*	15.67 *(2.57)*	5.27 *(1.43)*	4.25 *(1.08)*	31.75 *(3.60)*	3.75 *(1.34)*
Logistic	69.47 *(7.74)*	31.39 *(1.78)*	56.25 *(2.98)*	61.08 *(3.39)*	72.19 *(2.49)*	65.92 *(3.78)*	63.73 *(4.14)*	61.00 *(1.92)*	68.11 *(4.25)*
Neural nets	34.72 *(9.58)*	45.81 *(4.01)*	32.42 *(1.77)*	33.39 *(8.59)*	57.44 *(2.49)*	8.47 *(9.51)*	2.47 *(1.12)*	60.67 *(2.89)*	42.75 *(3.67)*
KNN	11.92 *(1.08)*	15.31 *(1.43)*	20.17 *(2.67)*	13.97 *(1.90)*	9.22 *(1.34)*	5.61 *(1.13)*	3.53 *(1.12)*	14.69 *(0.96)*	4.72 *(0.49)*
MP1	10.21 *(1.73)*	11.67 *(1.25)*	12.93 *(1.86)*	12.88 *(1.86)*	9.32 *(1.41)*	2.07 *(0.60)*	1.54 *(0.42)*	12.09 *(2.73)*	1.13 *(0.36)*
MP2	10.00 *(1.64)*	11.65 *(1.61)*	12.46 *(1.50)*	12.44 *(1.09)*	9.72 *(1.54)*	2.22 *(0.50)*	1.54 *(0.47)*	11.93 *(2.68)*	1.32 *(0.49)*
Integrated	6.53 *(1.01)*	10.23 *(3.45)*	12.92 *(1.56)*	6.13 *(0.57)*	4.76 *(0.70)*	2.56 *(0.65)*	2.05 *(0.96)*	8.01 *(1.52)*	1.97 *(0.58)*

Method	Sample 100	Sample 200	Sample 400	Sample 500
DA	13.00 *(9.55)*	24.50 *(5.78)*	18.77 *(3.75)*	18.81 *(3.15)*
Logistic	63.17 *(8.76)*	66.83 *(6.58)*	65.26 *(4.21)*	65.79 *(7.80)*
Neural nets	22.00 *(11.99)*	36.42 *(9.47)*	30.43 *(8.54)*	32.88 *(9.16)*
KNN	12.00 *(7.11)*	13.17 *(4.61)*	12.18 *(2.77)*	11.36 *(1.02)*
MP1	9.56 *(8.28)*	7.58 *(3.32)*	8.24 *(4.00)*	7.58 *(3.32)*
MP2	6.89 *(5.27)*	7.31 *(3.27)*	7.95 *(2.32)*	9.66 *(1.62)*
Integrated	6.70 *(2.84)*	6.80 *(1.59)*	6.57 *(1.35)*	6.22 *(0.95)*

[1]*The values in the brackets are standard deviations of the average number of misclassifications.*

For each data characteristic, the mean differences among the different methods were also compared. Table 5 shows the best performing methods based on multiple *t* statistics. The performances of almost all the methods are inferior to their base case, except for data characteristic such as: degree of correlation, unequal covariance, and outliers. This is

mainly due to the increased complexity in the problem situation. Therefore, more attention should be paid to relative performance change among the methods. The following discussion summarizes the observations from the results derived in this study.

Table 4: Misclassification Rates for the Validation Data (Hypotheses testing)[2]

Method	Base	Static/ Dynamic	Linearity assumption	Normality assumption	Outliers	Strong correlation	Unequal covariance	Sample proportion	Weak correlation
DA	21.42 *(2.75)*	31.75* *(3.36)*	35.21* *(2.93)*	37.75* *(5.63)*	16.63** *(1.92)*	7.41** *(1.57)*	4.83** *(1.45)*	37.21* *(4.51)*	4.13** *(1.74)*
Logistic	67.88 *(6.97)*	31.5* *(3.67)*	56.42* *(3.17)*	62.96 *(5.42)*	72.92 *(2.41)*	66.08 *(4.29)*	63.99 *(3.88)*	61.6* *(2.58)*	67.04 *(5.16)*
Neural nets	32.13 *(8.90)*	40.38* *(2.51)*	30.04 *(1.28)*	30.13 *(6.61)*	58.25* *(1.91)*	9.16** *(8.85)*	2.88** *(1.07)*	60.83* *(1.32)*	41.13* *(2.18)*
KNN	21.71 *(9.06)*	27.25* *(3.05)*	30.25* *(2.57)*	24.33* *(2.58)*	13.83** *(2.05)*	9.99** *(2.03)*	8.38** *(3.06)*	21.88 *(1.93)*	9.46** *(3.13)*
MP1	13.33 *(5.07)*	19.71* *(2.86)*	20.1* *(2.26)*	14.84 *(1.28)*	13.21 *(1.87)*	10.1** *(0.65)*	8.94** *(1.61)*	21.64* *(2.34)*	9.23** *(2.56)*
MP2	13.08 *(5.27)*	21.28* *(2.38)*	20.32* *(2.37)*	15.35* *(1.74)*	13.69 *(2.10)*	10.22** *(1.03)*	9.07** *(1.89)*	21.24* *(3.20)*	9.42** *(1.93)*
Integrated	10.87 *(4.59)*	21.71* *(3.44)*	17.05* *(2.03)*	13.22 *(3.03)*	8.05** *(1.09)*	8.98 *(3.26)*	3.26** *(1.00)*	14.35* *(3.64)*	2.55** *(1.25)*

Method	Sample 100	Sample 200	Sample 400	Sample 500
DA	24.75 *(19.88)*	31.25 *(2.28)*	25.48 *(6.63)*	20.45 *(2.95)*
Logistic	66.25 *(6.80)*	67.00 *(6.88)*	66.17 *(3.98)*	65.27 *(6.67)*
Neural nets	29.75 *(12.39)*	39.25 *(12.59)*	33.13 *(8.56)*	30.38 *(8.52)*
KNN	24.75 *(14.16)*	26.00 *(4.67)*	23.86 *(5.65)*	20.83 *(1.64)*
MP1	12.92 *(7.40)*	16.13 *(2.94)*	13.87 *(2.03)*	12.57 *(2.07)*
MP2	12.33 *(6.16)*	15.83 *(3.76)*	13.46 *(2.22)*	12.22 *(2.31)*
Integrated	14.50 *(7.97)*	13.11 *(3.40)*	12.69 *(3.71)*	10.46 *(1.49)*

[2]*The values in the brackets are standard deviations of the average number of misclassifications.*

* *Tests of significance, p<0.05, significantly higher than base case*

** *Tests of significance, p<0.05, significantly lower than base case*

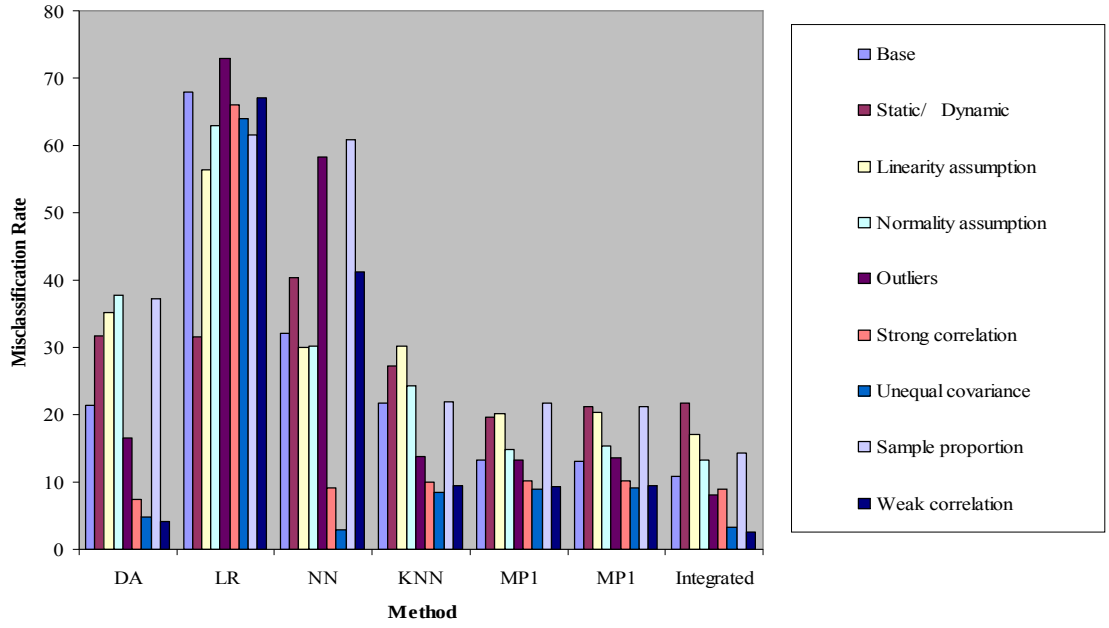

Figure 2: Effect of Data Characteristics on Validation Performance (Grouped by Method)

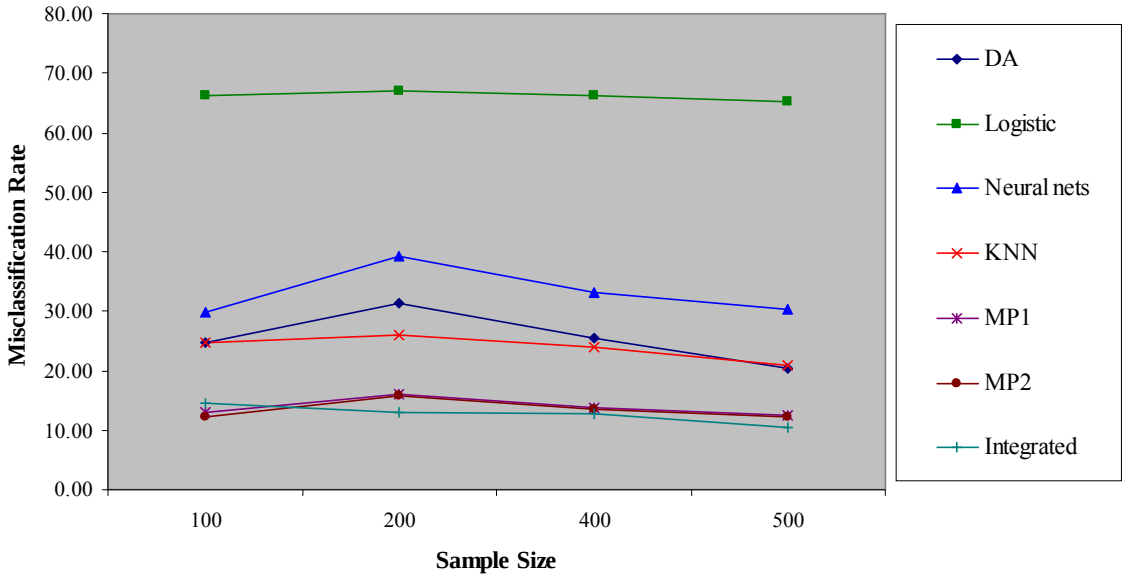

Figure 3: The Classification Performance versus the Sample Size

Table 5: Best Performing Methods under different Data Characteristics

Data characteristics	**Best performing methods**
Base case	Integrated, MP1, MP2
Dynamic scenario	Integrated, MP1, MP2
Nonlinearity	Integrated, MP1, MP2
Nonnormality	Integrated, MP1, MP2
Outliers	Integrated, MP1, MP2, KNN
Strong correlation	DA, Integrated, MP1, MP2, NN, KNN
Unequal covariance	NN, Integrated, DA
Unequal sample proportion	Integrated
Weak correlation	Integrated, DA
Sample size	Smaller sample: Integrated, MP1, MP2 Larger samples: Integrated, MP1, MP2

5.1 Analysis by method

In general, the performance of the mathematical programming techniques (LP1, LP2), and the integrated method are superior to DA, LR, NN, and k-NN for all data characteristics. Between the two mathematical programming techniques (LP1, LP2) and the integrated method, the integrated method significantly outperformed LP1 and LP2 under all the data characteristics. In some cases the performance of k-NN is comparable with LP1, and LP2. Furthermore, the variation in sample size has similar effect on almost all the methods and generally tends to agree with the previous research, i.e. as the sample size increases the classification accuracy increases. However, in this study, before an incremental decrease in misclassification rates for larger sample size, the misclassification rates hits a crest for a sample size of 200.

The performance of LR for all the data characteristics is inferior to all the other methods for both the training and validation data. The variation in sample sizes has little impact on its high misclassification rates, however, the misclassification rates decrease incrementally as the sample size increases.

The performance of DA under data characteristics such as strong and weak correlation, and unequal covariance structure is equal to or even superior to the LP1, LP2, and integrated methods. This is surprising considering the fact that DA is very sensitive to heterogeneity of variance-covariance matrices as well as multi-collinearity. In fact, one of the primary assumptions of DA is low multi-collinearity among predictor variables. However, its performance on other data characteristics such as dynamic environment, nonlinear data, nonnormal data, outliers, and unequal sample proportion is relatively lower, which are in line with the assumptions of DA. DA responds favorably to the increase in sample sizes which is indicated by a drastic dip in the misclassification rates for larger sample sizes.

NN performs well in strong correlation, and unequal covariance case, especially when the sample size is large. Infact, the performance of NN in unequal covariance case is superior to all the other methods. However, its performance on all the other data characteristics is inferior to all other methods except for LR. A comparison between the error rates shows that DA tends to do better than the NN when the level of non-linearity is low, but that neural networks do better when there is a greater degree of non-linearity.

Table 6: Performance of the integrated method versus the other methods (Hypotheses testing)[3]

Method	Integrated	DA	Logistic	Neural nets	KNN	MP1	MP2
Base	10.87	21.42*	67.88*	32.13*	21.71*	13.33*	13.08*
	(4.59)	*(2.75)*	*(6.97)*	*(8.90)*	*(9.06)*	*(5.07)*	*(5.27)*
Static/Dynamic	21.71	31.75*	31.5*	40.38*	27.25*	19.71	21.28
	(3.44)	*(3.36)*	*(3.67)*	*(2.51)*	*(3.05)*	*(2.86)*	*(2.38)*
Linearity assumption	17.05	35.21	56.42*	30.04	30.25*	20.1*	20.32*
	(2.03)	*(2.93)*	*(3.17)*	*(1.28)*	*(2.57)*	*(2.26)*	*(2.37)*
Normality assumption	13.22	37.75	62.96	30.13	24.33	14.84*	15.35*
	(3.03)	*(5.63)*	*(5.42)*	*(6.61)*	*(2.58)*	*(1.28)*	*(1.74)*
Outliers	8.05	16.63*	72.92*	58.25*	13.83*	13.21*	13.69*
	(1.09)	*(1.92)*	*(2.41)*	*(1.91)*	*(2.05)*	*(1.87)*	*(2.10)*
Strong correlation	8.98	7.41	66.08*	9.16	9.99	10.10	10.22
	(3.26)	*(1.57)*	*(4.29)*	*(8.85)*	*(2.03)*	*(0.65)*	*(1.03)*
Unequal covariance	3.26	4.83*	63.99*	2.88	8.38*	8.94*	9.07*
	(1.00)	*(1.45)*	*(3.88)*	*(1.07)*	*(3.06)*	*(1.61)*	*(1.89)*
Sample proportion	14.35	37.21*	61.6*	60.83*	21.88*	21.64*	21.24*
	(3.64)	*(4.51)*	*(2.58)*	*(1.32)*	*(1.93)*	*(2.34)*	*(3.20)*
Weak correlation	2.55	4.13	67.04*	41.13*	9.46*	9.23*	9.42*
	(1.25)	*(1.74)*	*(5.16)*	*(2.18)*	*(3.13)*	*(2.56)*	*(1.93)*

Method	Sample 100	Sample 200	Sample 400	Sample 500
Integrated	14.50	13.11	12.69	10.46
	(7.97)	*(3.40)*	*(3.71)*	*(1.49)*
DA	24.75	31.25*	25.48	20.45*
	(19.88)	*(2.28)*	*(6.63)*	*(2.95)*
Logistic	66.25*	67.01*	66.17*	65.27*
	(6.80)	*(6.88)*	*(3.98)*	*(6.67)*
Neural nets	29.75*	39.25*	33.12*	30.38*
	(12.39)	*(12.59)*	*(8.56)*	*(8.52)*
KNN	24.75*	26.01*	23.86*	20.83*
	(14.16)	*(4.67)*	*(5.65)*	*(1.64)*
MP1	12.92	16.12	13.87	12.57*
	(7.40)	*(2.94)*	*(2.03)*	*(2.07)*
MP2	12.33	15.83	13.46	12.22
	(6.16)	*(3.76)*	*(2.22)*	*(2.31)*

[3]*The values in the brackets are standard deviations of the average number of misclassifications.*

* *Tests of significance, $p<0.05$, significantly higher than base (Reject H_o at $\alpha=0.05$).*

This is in line with the research findings by Curram and Mingers (1994), which presents an empirical comparison of three classification methods: neural networks, decision tree

induction and linear discriminant analysis, based on seven datasets with different characteristics, four being real, and three artificially created.

The performance of k-NN for most of the data characteristics is comparable with both the mathematical programming methods, i.e. LP1 and LP2. However, in case of dynamic environment, nonlinearity and nonnormality assumptions, LP1 and LP2 perform relatively better than the k-NN. Furthermore, k-NN significantly outperformed DA when normality and linearity assumptions are not in place. k-NN is also superior to DA in presence of outliers and unequal sample proportion cases. However, DA did better in the unequal covariance case. This result is somewhat contrary to the work by Kiang (2003), where DA performs better than k-NN in the unequal covariance case. k-NN shows a gradual decline in misclassification rates as the sample size increases. Another feature of k-NN which needs to be mentioned here is that its classification performance in case of a binary categorical variable is superior to its performance in the present case of a categorical variable with four ordinal classes (Dreiseitl et al. 2001, Kiang 2003). This indicates that an increase in the problem complexity deteriorates the classification performance of k-NN (Shmueli et al. 2006).

The performances of both the mathematical programming approaches, i.e. LP1 and LP2 are superior compared with all other methods except for the integrated method, on most data characteristics. There is empirical evidence that these nonparametric methods may produce more accurate classification rules than the traditional statistical methods (Gochet et al. 1997). As opposed to parametric approaches, mathematical programming

approaches allow individual observations to be weighted, or relative penalties for misclassification to be set, providing more flexibility to decision-makers (Freed and Glover 1981b, Koehler 1990, Erenguc and Koehler 1990). Between the two methods, LP1 and LP2, there is not much difference in their performances on all the data characteristics. The LP2 method with its emphasis on median minimization was expected to perform better than LP1 on the outlier data, surprisingly; however, their misclassification rates are not significantly different from each other. Both the methods show a decrease in misclassification rates as the sample size increases.

The performance of the proposed integrated (hybrid) method is superior to almost all the other methods for all data characteristics and sample sizes on both training and validation data (see Table 6). In most cases, we reject the null hypothesis at 5 % significance level. This concurs with our hypothesis. However, when the data characteristic is dynamic, its performance does not differ considerably from the MP1 and MP2, and when the data is strongly correlated, DA performs slightly better than the integrated method. Furthermore, its relatively lower standard deviations show that the method is fairly robust. The results indicate that the integrated method utilizes the strengths of both, k-NN and LP approach and performs better than the usual k-NN or LP for almost all the data characteristics.

5.2 Analysis by data characteristics

The dynamic environment affects the relative classification accuracy of almost all the methods. However, MP1, MP2, and the integrated method are moderately affected by it.

Surprisingly, the performance of the neural network model is relatively inferior, given that they have been found to handle both dynamic and static problems well due to its ability of respond swiftly to changes in the real world.

The linearity assumption significantly affects the performance of all the methods except the neural network, compared to their respective base case. The neural network is not significantly affected by the linearity assumption, a reason being, neural networks allow nonlinear relations and complex interactions among predictor variables and thus score over parametric methods (Kotsiantis et al. 2006, Paliwal and Kumar 2009). Like linearity, the normality assumption significantly affects the performance of DA, which understandably is due to the violation of one of the fundamental assumptions of the statistical technique.

The performances of DA, k-NN, and the integrated method are superior to their respective base cases, in the presence of the outliers. The performance of the neural network is significantly affected by the presence of outliers, which concurs with our hypothesis in Table 1. Khamis et al. (2005) carried out a study to investigate the influence of outliers on neural network performance in two ways; by examining the percentage outliers and secondly the magnitude outliers. The authors conclude that both: the percentage outliers and magnitude outliers affect the neural network performance. Surprisingly, the performance of MP2, a median minimization LP, specifically designed to handle the outliers, is slightly lower than MP1.

Almost all the methods except for LR show a significant improvement in classification accuracy in the presence of multicollinearity among the predictor variables. In fact, the performance of DA is superior in the presence of multicollinearity, which is contrary to one of its assumptions. Past research reveals that the NN performs well when multicollinearity is present and a nonlinear relationship exists between the input and the output features (Kotsiantis et al. 2006). This is further reinforced by its relatively poor performance in the absence of multicollinearity. Again, all methods except LR and neural network show a significant improvement in performance in the absence of multicollinearity.

Sample size has a significant effect on DA and NN methods. For all the other methods, the decrease in misclassification is rather incremental as the sample size increases (see Figure 3). A peculiar observation of the effect of sample size is that the performance of all methods deteriorate when the sample size is 200, however, the performance shows an increasing trend as the sample size increases. In Table 6, we test the effect of sample size on various methods compared with our integrated method. Only, MP1 and MP2 methods compare favorably with the performance of integrated method, whereas, LR, NN and k-NN methods have significantly higher misclassification rates. Our results concur with the extant research on the topic of effect of sample sizes on classification methods (Raudys and Jain 1991, Ho 1998, Kiang 2003, Maas and Hox 2005).

The unequal covariance bias has a significant impact on all the methods except LR, and helps in reducing the misclassification rates. On the contrary, the unequal sample

proportion bias significantly reduces the classification accuracy of all the methods except k-NN.

Based on our finding and the analysis of results, we can adjust and update Table 1. The revised version of Table 1 is shown in Table 7.

5.3 Analysis by individual error rates

The results of training and validation data for individual error rates are shown in Table 8 and Table 9, respectively. We judge the effectiveness of a method on individual error rates by its ability to accurately classify top customer groups, i.e. Y = 1, and Y = 2. The results show that individual error rates in case of almost all methods are affected by the data characteristics. However, LR is the most affected, whereas the mathematical programming approaches (LP1, and LP2) including integrated method are the least affected for both training and validation data.

In general, the LP1, LP2, and integrated methods have relatively higher and stable classification accuracy under all the data characteristics for the top customer groups. On the other hand, DA and k-NN have higher classification accuracy for group Y = 1, i.e. prime customers but their performance is erratic for group Y = 2, i.e. high value customers. Another concern, which is exhibited by the results, is that all the methods perform relatively poorly in classifying Group 0 (i.e. 'No Buyer' customer class).

Table 7: Revised version of Table 1 (Hypotheses)

Method	Hypotheses	Experimental results
DA	Static scenario. Affected by normality and linearity violations, low correlations, outliers and identical covariances.	Static scenario. Affected by linearity and normality violations, as well as unequa sample proportion. Strong and weak correlation, Presence of outliers and unequa covariance structure improves performance.
Logistic	Static scenario. Affected by sample size, especially when dependent variable has many groups, and low correlations.	Static scenario. Affected by nonlinearity and unequal sample proportion. Worst performing method.
Neural nets	Both static and dynamic scenarios. Affected by sample size and outliers.	Affected by dynamic scenario, outliers, unequal sample proportion, and weak correlation. Strong correlation and unequal covariance structure improves performance.
KNN	Static scenario. Affected by sample size and outliers.	Static scenario. Affected by nonlinearity, nonnormality and sample size, but unaffected by unequal sample proportion. Presence of outliers, strong and weak correlation and unequal covariance structure improve performance.
MP1	Static scenario. Affected by linearity violations.	Static scenario. Affected by nonlinearity and unequal sample proportion, but unaffected by nonnormality and outliers. Degree of correlation and unequal covariance improves performance.
MP2	Static scenario. Affected by linearity violations.	Static scenario. Affected by nonlinearity, nonnormality and unequal sample proportion, but unaffected by presence of outliers. Degree of correlation and unequal covariance improves performance.
Integrated	Static scenario. Affected by linearity violations and sample size.	Static scenario. Affected by nonlinearity and unequal sample proportion, but unaffected by nonnormality and multicollinearity. Presence of outliers and

The performances of all method except LR are superior under the data characteristics such as the degree of correlation (strong and weak), and unequal covariance compared with the base case. The data characteristics such as the dynamic environment, nonlinearity, nonnormality, and unequal sample proportion adversely affects individual error rate for almost all methods. However, NN is relatively less affected by the nonlinearity of data. Furthermore, though NN allows adaptive model adjustments and responds swiftly to changes in the real world, its performance on the dynamic environment is mixed, in that, its classification accuracy for group, Y = 2 is high, however, its accuracy for Y = 1 is low.

Table 10 shows the effect of sample sizes on the individual error rates. The results show that the integrated method gives superior performance for both the groups of interest, i.e. Group 1 (i.e. prime customers) and Group 2 (i.e. high value customers). There is an incremental improvement in its performance with the increase in sample size. The DA, k-NN algorithm, and the linear programming approaches (LP1 and LP2) have relatively lower individual error rates for Group 1; however, their performance deteriorates while classifying Group 2.

Overall, the results indicate that the data characteristic does affect the individual group error rates for all the methods. Companies with large amounts of customer data pay considerable attention to the analysis of data to target appropriate customer segments for their products and services. Database marketing uses the power of data and information

Table 8: Individual Error Rates for Training Data

Methods	Groups	Base case*	Dynamic	Nonlinearity	Nonnormality	Outliers	Strong correlation	Unequal covar	Sample proportion	Weak correlation
DA*	0	*31.13%	46%	58%	38%	29%	8%	5%	26%	5%
	1	1%	42%	0%	2%	3%	3%	8%	8%	6%
	2	11%	34%	74%	23%	8%	6%	4%	12%	4%
	3	37%	2%	4%	73%	24%	4%	0%	64%	0%
Logistic	0	54%	55%	92%	99%	57%	30%	22%	6%	5%
	1	43%	60%	94%	99%	94%	71%	76%	100%	88%
	2	96%	10%	29%	29%	87%	97%	93%	100%	100%
	3	86%	1%	10%	18%	50%	66%	75%	95%	92%
Neural nets	0	77%	86%	100%	100%	99%	7%	3%	100%	6%
	1	48%	85%	0%	8%	100%	7%	3%	9%	100%
	2	8%	4%	24%	19%	31%	9%	4%	99%	62%
	3	0%	0%	4%	0%	0%	11%	0%	0%	0%
KNN	0	16%	14%	25%	15%	8%	4%	2%	11%	2%
	1	0%	13%	1%	1%	1%	1%	3%	2%	2%
	2	20%	32%	42%	23%	15%	6%	4%	28%	9%
	3	12%	3%	15%	20%	12%	12%	6%	15%	5%
MP1	0	18%	19%	20%	24%	16%	2%	1%	14%	0%
	1	2%	18%	2%	1%	2%	1%	1%	3%	0%
	2	7%	8%	18%	12%	8%	3%	1%	9%	1%
	3	13%	2%	12%	15%	12%	3%	3%	17%	3%
MP2	0	18%	19%	19%	23%	17%	2%	1%	13%	0%
	1	1%	18%	2%	1%	2%	1%	1%	3%	1%
	2	7%	8%	18%	11%	8%	3%	2%	9%	1%
	3	13%	2%	11%	14%	12%	4%	3%	16%	3%
Integrated	0	26%	22%	38%	22%	13%	7%	4%	16%	4%
	1	3%	10%	6%	3%	3%	1%	2%	6%	2%
	2	3%	10%	10%	3%	3%	2%	2%	6%	2%
	3	3%	3%	9%	3%	3%	2%	2%	5%	1%

* Reading the table: Under base case, for DA, 31% of Group 0 observations have been misclassified into Groups 1, 2, and 3.

Table 9: Individual Error Rates for Validation Data

Methods	Groups	Base case	Dynamic	Nonlinearity	Nonnormality	Outliers	Strong correlation	Unequal covar	Sample proportion	Weak correlation
DA	0	37%	48%	60%	42%	29%	11%	7%	28%	5%
	1	3%	43%	0%	3%	3%	4%	5%	11%	7%
	2	9%	36%	76%	24%	10%	9%	6%	14%	5%
	3	37%	3%	3%	75%	23%	6%	1%	67%	0%
Logistic	0	49%	54%	94%	99%	61%	28%	22%	6%	4%
	1	43%	60%	93%	99%	93%	70%	75%	100%	85%
	2	95%	11%	27%	31%	85%	99%	95%	100%	100%
	3	85%	2%	12%	23%	53%	68%	75%	97%	92%
Neural nets	0	79%	88%	100%	100%	99%	8%	4%	100%	6%
	1	50%	82%	0%	9%	100%	8%	2%	12%	100%
	2	8%	4%	20%	20%	34%	10%	3%	98%	63%
	3	0%	1%	3%	0%	0%	10%	2%	0%	0%
KNN	0	33%	26%	42%	30%	16%	9%	5%	19%	6%
	1	1%	26%	3%	1%	5%	3%	5%	7%	6%
	2	31%	49%	55%	33%	21%	10%	11%	39%	18%
	3	21%	7%	19%	30%	15%	18%	12%	19%	10%
MP1	0	38%	47%	49%	30%	41%	34%	31%	30%	28%
	1	2%	19%	2%	0%	2%	0%	1%	3%	1%
	2	8%	6%	24%	14%	7%	3%	2%	11%	3%
	3	6%	7%	5%	15%	3%	3%	2%	27%	5%
MP2	0	34%	54%	49%	24%	39%	32%	30%	28%	27%
	1	1%	19%	2%	1%	2%	1%	2%	4%	1%
	2	7%	7%	24%	14%	7%	4%	2%	11%	3%
	3	9%	5%	5%	22%	7%	4%	3%	28%	6%
Integrated	0	36%	32%	60%	45%	23%	11%	8%	27%	7%
	1	6%	33%	11%	5%	5%	3%	3%	14%	2%
	2	6%	18%	7%	11%	5%	10%	2%	10%	2%
	3	6%	8%	9%	5%	5%	13%	2%	9%	1%

technology in the pursuit of personal marketing of products and services to consumers, based on their preferences and needs (Zahavi and Levin 1997b). The importance of individual group error rates analysis can be judged from this fact. This analysis should help the practioners to understand the relative importance of various methods vis-à-vis different data characteristics and choose a method that best helps in identifying their target segments.

Table 10: Individual Error Rates for Sample Size for Training and Validation Data

Methods	Groups	Sample 100	Sample 200	Sample 400	Sample 500
DA	0	38%	52%	41%	32%
	1	2%	2%	2%	1%
	2	6%	13%	9%	10%
	3	11%	33%	26%	34%
Logistic	0	67%	75%	66%	56%
	1	85%	90%	71%	37%
	2	63%	65%	76%	99%
	3	37%	37%	54%	86%
Neural nets	0	65%	83%	74%	73%
	1	21%	19%	28%	45%
	2	6%	37%	17%	8%
	3	1%	9%	3%	0%
KNN	0	19%	18%	18%	17%
	1	2%	0%	1%	0%
	2	15%	29%	22%	22%
	3	14%	9%	11%	12%
MP1	0	18%	14%	16%	16%
	1	1%	1%	1%	2%
	2	12%	5%	8%	8%
	3	8%	10%	10%	14%
MP2	0	13%	12%	14%	18%
	1	4%	1%	2%	1%
	2	5%	6%	6%	8%
	3	5%	10%	9%	12%
Integrated	0	23%	27%	25%	24%
	1	2%	2%	2%	3%
	2	5%	4%	4%	3%
	3	7%	5%	5%	3%

Methods	Groups	Sample 100	Sample 200	Sample 400	Sample 500
DA	0	48%	63%	50%	39%
	1	5%	3%	3%	3%
	2	9%	13%	11%	9%
	3	26%	42%	35%	35%
Logistic	0	76%	77%	67%	49%
	1	86%	92%	74%	45%
	2	65%	63%	73%	90%
	3	38%	37%	53%	83%
Neural nets	0	72%	87%	75%	68%
	1	29%	19%	32%	47%
	2	11%	36%	19%	9%
	3	1%	13%	5%	0%
KNN	0	37%	41%	38%	36%
	1	6%	1%	3%	2%
	2	25%	37%	30%	29%
	3	25%	19%	22%	22%
MP1	0	23%	32%	30%	37%
	1	1%	2%	2%	2%
	2	21%	17%	15%	7%
	3	7%	14%	9%	6%
MP2	0	22%	31%	30%	37%
	1	6%	3%	3%	1%
	2	16%	17%	13%	7%
	3	6%	12%	9%	8%
Integrated	0	50%	45%	44%	37%
	1	4%	8%	6%	6%
	2	13%	7%	8%	5%
	3	6%	10%	7%	6%

CHAPTER 6: CONCLUSIONS

In this computational experimental study we have compared seven different methods of classification: discriminant analysis – Mahanalobis (DA), multinomial logistic regression(LR), neural network (NN), k-nearest neighbor algorithm (kNN), two variants of linear programming (MP1, and MP2) and, an integrated (hybrid) method, in different settings with respect to the distributions of the discriminating variables, the correlation structures between the variables and the absence or presence of outliers in the data set, unequal covariance among various groups, and a dynamic environment. We have used four different groups and the classification errors of each of the methods, and their individual error rates, on each of the data characteristics are recorded.

Using the characteristics of a real data set, Monte Carlo simulation experiments were used to generate multiple data sets for each of the data characteristics mentioned above. Based on previous research, we use the apparent error rate as a performance evaluation measure. The controlled experiments conducted show that the classification algorithms are sensitive to changes in data characteristics. The misclassification rates dues to biases can be substantially high in the presence of even a single bias, as seen in the case of dynamic environment.

The study shows that the proposed integrated method, which is a combination of k-NN algorithm and a linear programming approach, dominates almost all the other methods on the classification performance. Moreover, its performance is better than the k-NN and

LP approach, individually, an indicator of the utilization of the strengths of both the methods for improved classification accuracy. Logistic regression and neural network methods provide worst relative performance under most scenarios. This result contradicts some of the previous research studies and reviews (Dreiseitl and Ohno-Machado 2002, Kiang 2003, Paliwal and Kumar 2009). Multinomial logistic regression is a parametric method for prediction and classification but its performance depends on the distribution of variables, size and quality of data (Sadat-Hashemi et al. 2004). This study clearly establishes that the data complexities such as: multicollinearity, heterogeneity and nonlinear relations among response and predictors have adverse impact on the multinomial logistic regression model. The fluctuation in the Neural network model's performance can be attributed to the large number of possible parameter settings and the absence of a methodical approach to choosing the best settings. For example, experiments must be conducted to determine the best data representation, model specification, number of hidden layers, number of neurons on each hidden layer, learning rate, and number of training cycles. All of these interrelate to give the best ANN model. Failure to conduct such experiments may result in a poorly specified ANN model (Nguyen and Cripps 2001).

The performance of the linear programming methods such as MP1 and MP2 does not lag far behind the superior performance of the integrated methods. In fact, their performances are better than statistical methods, neural network, and k-NN under data characteristics such as dynamic environment, nonlinearity, nonnormality, unequal sample proportion and in the presence of outliers. The only problem with this nonparametric

method is the computational time required for the execution. However, with the advent of faster and powerful computing machines this glitch should not pose much problem in its utility as a robust and relative accurate classifier.

An important concern brought forth by our results is the impact of dynamic variations in data and unequal sample proportion on classification performance. The results indicate that all classification methods including the integrated method are adversely affected by the nonstatic nature of the data. Since most business phenomenon exhibit dynamic behavior, care should be exercised in calibrating classification systems to such scenarios.

Overall, an important result of this study is the demonstration of the effectiveness of the integrated method in improving the classification accuracy on both training and validation data for most of the data circumstances. Furthermore, the importance of linear programming approaches to achieve the goal of improved classification also needs to be highlighted.

CHAPTER 7: LIMITATIONS AND FUTURE RESEARCH

This study has shown the effectiveness of an integrated (hybrid) method for all methods and under various data circumstances. The study also conclusively proves the efficacy of linear programming approaches. However, there is further scope for exploring various hybrid techniques that combines the strengths of different methods to improve classification accuracy. Though, with the advances in computing technology, the time and efforts taken by linear programming approaches can be overlooked, but this is still a factor which could limit the use of these techniques by practioners.

In this study, we study a financial problem with three predictor variables. Further research involving more attributes could help gain more insights into the relative strengths of the methods. Another area for further investigation could be including more observations in the problem as well as varying the training and validation data sets, to test the robustness of the methods.

For each data characteristic, several versions of biases should be used to test the models. For example, our study uses a lognormal distribution to introduce nonnormality biases into the data. Future research may use other types of distributions such as exponential, uniform, etc., to gain full understanding of the impact of these factors.

More sophisticated experiments are required to examine the possible interactions among the predictors on various methods. However, this may pose a serious challenge to linear

programming approaches, k-NN and neural network, in terms of the problem complexities.

The study compared only seven different methods. Future work could include more methods such as decision tree (C4.5), different variations of neural network, support vector machines (SVMs), and others including few hybrid methods.

CHAPTER 8: RESEARCH CONTRIBUTIONS

Building on the previously cited work, the present study contributes new knowledge to several areas of multi-group classification. Some of these contributions are:

1. Previous researches show, few studies have touched upon more than three group classification. Our study delves comprehensively on a four-group classification problem using several classification techniques such as discriminant analysis (Mahanalobis distance), multinomial logistic regression, neural network, k-NN algorithm, two variants of LP, and a proposed integrated method.

2. The performance of our integrated method in terms of its classification accuracy and lower individual group error rates under robust experimental conditions shows the utility of an integrated (hybrid) technique, especially its ability to combine the strengths of k-NN and linear programming approach.

3. Very few previous researches have studied the individual group error rates of each of the methods under such varied data circumstances. Our study calculates the individual group error rates for each of the methods, which provides tremendous insight to the practioners in their choice of methods.

4. Our study provides new insights on the efficacy of linear programming methods vis-à-vis statistical techniques such as discriminant analysis, multinomial logistic

regression, neural network, and k-NN. Very few studies have demonstrated the behavior of all of above methods under so many data characteristics as this study has done. Furthermore, most previous such studies have been confined to two-group classification problems.

REFERENCES

Allenby, G., P. Lenk. 1994. Modeling Household Purchase Behavior with Logistic Normal Regression. *Journal of the American Statistical Association*. 89(428) 1218-1231.

Altman, E. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*. 23(4) 589-609.

Altman, E., G. Marco, F. Varetto. 1994. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance*. 18 505–529.

Anderson, J., R. Narasimhan. 1979. Assessing Project Implementation Risk: A Methodological Approach. *Management Science*. 25(6) 512-521.

Asparoukhova, O., W. Krzanowskib. 2001. A comparison of discriminant procedures for binary variables. *Computational Statistics & Data Analysis*. 38 139–160.

Awh, R. Y., D. Waters. 1974. A discriminant analysis of economic, demographic, and attitudinal characteristics of bank charge-card holders: A case study. *Journal of Finance*. 29 973-983.

Bajgier, S., A. Hill. 1982. An experimental comparison of statistical and linear programming approaches to the discriminant problems. *Decision Sciences*. 13 604-618.

Bal, H., H. Orkcu, S. Celebioglu. 2006. An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems. *Computers & Industrial Engineering*. 50 296–311.

Ball, C., A. Tschoegl. 1982. The Decision to Establish a Foreign Bank Branch or Subsidiary: An Application of Binary Classification Procedures. *The Journal of Financial and Quantitative Analysis*. 17(3) 411-424.

Bentz, Y., D. Merunka. 2000. Neural networks and the multinomial logit for brand choice modeling: A hybrid approach. *Journal of Forecasting*. 19 177-200.

Bloom, J. 2004. Tourist market segmentation with linear and non-linear techniques. *Tourism Management*. 25 723–733.

Bloom, J. 2005. Market segmentation - A neural network application. *Annals of Tourism Research*. 32(1) 93–111.

Boone, D., M. Roehm. 2002. Evaluating the appropriateness of market segmentation solutions using artificial neural networks and the membership clustering criterion. *Marketing Letters*. 13(4) 317–333.

Breiman, L. 1996a. Stacked regression. *Machine Learning*. 24 49– 64.

Breiman, L. 1996b. Bagging predictors. *Machine Learning*. 24 123– 140.

Brockett, P., W Cooper, L Golden, X. Xia. 1997. A case study in applying neural networks to predicting insolvency for property and casualty insurers. *Journal of the Operational Research Society*. 48(12) 1153-1162.

Buttrey, S., C. Karo. 2002. Using k-nearest-neighbor classification in the leaves of a tree. *Computational Statistics and Data Analysis*. 40 27–37.

Capon, N. 1982. Credit scoring systems: A critical analysis. *Journal of Marketing*. 46 82-91.

Carvalho, D., F. Alex. 2004. A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences*. 163 13–35.

Chen, P. 2003. A hybrid framework using SOM and fuzzy theory for textual classification in data mining. *Modeling with Words*. LNAI2873 153–167.

Cheng, B., D Titterington. 1994. Neural networks: A review from a statistical perspective. *Statistical Science*. 9(1) 2-30.

Coenen, F., Swinnen, G., Vanhoof, K., Wets, G. 2000. The improvement of response modeling: combining rule-induction and case-based reasoning. *Expert Systems with Application*. 18(4) 307–313.

Curram, S., J. Mingers. 1994. Neural Networks, Decision Tree Induction and Discriminant Analysis: an Empirical Comparison. *The Journal of Operational Research* Society. 45(4) 440-450.

Dasgupta, G., G. Dispensa, S. Ghose. 1994. Comparing the predictive performance of a neural network model with some traditional market response models. *International Journal of Forecasting*. 10(2) 235-244.

Dasarathy, B. 1973. An integrated non-parametric sequential approach to multi-class pattern classification. *International Journal of Systems Science*. 4 449-457.

Dasarathy, B., B. Sheela. 1979. A composite classifier system design: concepts and methodology. *Proceedings of the IEEE*. 67(5) 708-713.

Devijver, P., J. Kittler. 1982. Pattern Recognition: A Statistical Approach. *Prentice Hall International*, London.

Dillon, W., R. Calantone, P. Worthing. 1979. The New Product Problem: An Approach for Investigating Product Failures. *Management Science*. 25(12) 1184-1196.

Doumpos, M., C. Zopounidis. 2002. Multicriteria Decision Aid Classification Methods. *Kluwer Academic Publishers, Inc*., Boston, MA.

Dreiseitl, S., L. Ohno-Machado. 2002. Logistic regression and artificial neural network classificationmodels: a methodology review. *Journal of Biomedical Informatics*. 35 352–359.

Dreiseitl, S., L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, M. Binder. 2001. A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions. *Journal of Biomedical Informatics*. 34 28–36.

Duda, R., E. Hart. 1973. Pattern classification and scene analysis. *Wiley*, New York, NY.

Dutta, S., S. Shekhar. 1988. Bond Rating: A non-conservative application of neural network. *IEEE International Conference on Neural Networks*. 2 443-450.

Dyer, A. 1974. Comparisons of tests for normality with a cautionary note. *Biometrika*. 61(1) 185-189.

Eisenbeis, R. 1977. Pitfalls in the application of discriminant analysis in business, finance, and economics. *Journal of Finance*. 32(3) 875.

Erenguc, S., G Koehler 1990. Linear Programming Methods for Discriminant Analysis: Introduction. *Managerial and Decision Economics*. 11(4) 213-214.

Fausett, L. 1994. Fundamentals of Neural Networks. Prentice Hall, Inc., Upper Saddle River, NJ.

Fish, K., J. Barnes, M. Aiken. 1995. Artificial neural networks - A new methodology for industrial market segmentation. *Industrial Marketing Management*. 24 431-438.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. 7 179–188.

Flitman, A. 1997. Towards analyzing student failures: Neural networks compared with regression analysis and multiple discriminant analysis. *Computers & Operations Research*. 24(4) 367-377.

Freed, N., F. Glover. 1981a. A linear programming approach to the discriminant problem. *Decision Sciences*. 12 68-74.

Freed, N., F. Glover. 1981b. Simple but powerful goal programming models for discriminant problems. *European Journal of Operational Research*. 7 44-60.

Freed, N., F. Glover. 1982. Linear programming and statistical discrimination – The LP side. *Decision Sciences*. 13 172-175.

Freed, N., F. Glover. 1986a. Evaluating alternative linear programming models to solve the two-group discriminant problem. *Decision Sciences*. 17 151-162.

Freed, N., F. Glover. 1986b. Resolving certain difficulties and improving the classification power of LP discriminant analysis formulations. *Decision Sciences*. 17 589–595.

Fukunaga, K., T. Flick. 1984. An optimal global nearest neighbor metric. *IEEE Transaction*. *Pattern Analysis*. PAMI-6 3 314–318.

Galguera, L., D. Luna, M. Méndez. 2006. Predictive segmentation in action: Using CHAID to segment loyalty card holders. *International Journal of Market Research*. 48 (4) 459-479.

Gallant, S. 1988. Connectionist expert systems. *Communications of the ACM*. 31(2) 152 - 169.

Gan, C., V. Limsombunchai, M. Clemes, A. Weng. 2005. Consumer Choice Prediction: Artificial Neural Networks versus Logistic Models. *Journal of Social Sciences*. 1(4) 211-219.

Gehrlein, W. 1986. General Mathematical Programming Formulations for the Statistical Classification Problem. *Operations Research Letters*. 5 299-304.

Glorfeld, L., N. Gaither. 1982. On using linear programming in discriminant problems. *Decision Sciences*. 13 167-171.

Glorfeld, L., M. Kattan. 1989. A comparison of the performance of three classiFcation procedures when applied to contaminated data. *Proceedings of the 21st Annual Meeting of the Decision Sciences Institute*. New Orleans, Louisiana. 1153–1155.

Glover, F. 1990. Improved linear programming models for discriminant analysis. *Decision Sciences*. 21 771–785.

Glover, F., S. Keene, B. Duea. 1988. A new class of models for the discriminant problem. Decision Sciences. 19 269–280.

Gochet, W., A. Stam, V. Srinivasan, S. Chen. 1997. Multi-group Discriminant Analysis Using Linear Programming. *Operations Research*. 45(2) 213-225.

Gyan, B., K Voges, N. Pope. 2005. Artificial Neural Networks in Marketing from 1999 to 2003: A Region of Origin and Topic Area Analysis. *Proceedings of the ANZMAC 2005 Conference*. Wellington, New Zealand.

Hample, F., E. Ronchetti, P. Rousseeuw, W. Stahel. 1986. Robust statistics: the approach based on influence functions. *Wiley*, New York, NY.

Hand, D. 2001. Modelling consumer credit risk. *IMA Journal of Management Mathematics*. 12 139-155.

Hansen, L. K., Salaman, P. 1990. Neural networks ensembles. *Transactions on Pattern Analysis and Machine Intelligence*. 12(10) 993–1001.

Haykin, S. 1999. Neural networks: A comprehensive foundation. Second Edition. *Prentice-Hall Inc*., Upper Saddle River, NJ.

Hertz, J., R Palmer, A. Krogh. 1991. Introduction to the Theory of Neural Computation. *Westview Press*, Boulder, CO.

Ho, T. 1998. Nearest neighbors in random subspaces. *Lecture Notes in Computer Science*. 1451 640-648. Springer Berlin, Germany.

Hoaglin, D. 1985. Summarizing shape numerically: The *g*-and-*h* distributions, in: D. Hoaglin, F. Mosteller, J. Tukey (eds.). Exploring data, tables, trends, and shapes. *Wiley*, NY.

Hosmer, D., S. Lemeshow. 2001. Applied Logistic Regression. *John Wiley and Sons Inc*., NJ.

Hosseini, J., R. Armacost. 1994. The two-group discriminant problem with equal group mean vectors: An experimental evaluation of six linear/nonlinear programming formulations. *European Journal of Operational Research*. 77 241-252.

Hruschka, H., M. Natter. 1999. Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research*. 114 346-353.

Hur, J., J. Kim. 2008. A hybrid classification method using error pattern modeling. *Expert Systems with Applications*. 34(1) 231–241.

Indurkhya, N., S. Weiss. 1998. Estimating performance gains for voted decision trees. *Intelligent Data Analysis*. 2(4) 303–310.

Joachimsthaler, E., A. Stam. 1988. Four approaches to the classification problem in discriminant analysis: An experimental study. *Decision Sciences*. 19 322–333.

Joachimsthaler, E., A. Stam. 1990. Mathematical programming approaches for the classification problem in two-group discriminant problem. *Multivariate Behavioral Research*. 25 427-457.

Kaefer, F., C. Heilman, S. Ramenofsky. 2005. A neural network application to consumer classification to improve the timing of direct marketing activities. *Computers & Operations Research*. 32 2595–2615.

Kanal, L., B. Chandrasekaran. 1972. On linguistic, statistical and mixed models for pattern recognition in Frontiers of Pattern Recognition. S. Watanabe (Eds.). Academic Press, New York, NY.

Kartalopoulos, S. 1996. Understanding neural networks and fuzzy logic. *IEEE Press*. Piscataway, NJ.

Khamis, A., Z. Ismail, K. Haron, A. Tarmizi Mohammed. 2005. The Effects of Outliers Data on Neural Network Performance. *Journal of Applied Science*. 5(8) 1394-1398.

Kiang, M. 2003. A comparative assessment of classification methods. *Decision Support Systems*. 35 441– 454.

Kim, J., S. Wei, H. Ruys. 2003. Segmenting the market of West Australian senior tourists using an artificial neural network. *Tourism Management*. 24 25–34.

Koehler, G. 1990. Considerations for mathematical programming models in discriminant analysis. *Managerial and Decision Economics*. 11 227–234.

Kotsiantis, S., D. Kanellopoulos, V. Tampakas. 2006. On Implementing a Financial Decision Support System. *International Journal of Computer Science and Network Security*. 6(1A) 103-112.

Kumar, A., V. Rao, H. Soni. 1995. An Empirical Comparison of Neural Network and Logistic Regression Models. *Marketing Letters*. 6(4) 251-263.

Kuncheva, I., C. Bezdek, M. Shutton. 1998. On combining multiple classifiers by fuzzy templates. *International conference on artificial neural networks, IEEE*. 193–197.

Kwak, N., S. Kim, C. Lee, T. Choi. 2002. An Application of Linear Programming Discriminant Analysis to Classifying and Predicting the Symptomatic Status of HIV/AIDS Patients. *Journal of Medical Systems*. 26(5) 427-438.

Lacher, R., P. Coats, S. Sharma, L. Fant. 1995. A neural network for classifying the financial health of a firm. *European Journal of Operational Research*. 85 53-65.

Lam, K., J. Moy. 2003. A simple weighting scheme for classification in two-group discriminant problems. *Computers & Operations Research*. 30 155–164.

Lam, K., E. Choo, J. Moy. 1996a. Minimizing deviations from the group mean: A new linear programming approach for the two-group classification problem. *European Journal of Operational Research*. 88 358-367.

Lam, K., E. Choo, J. Moy. 1996b. Improved Linear Programming Formulations for the Multi-Group Discriminant Problem. *Journal of the Operational Research Society*. 47(12) 1526-1529.

Lawrence, K., D. Pai, R. Klimberg, S. Kudyba, S. Lawrence. 2007. A Classification Model for Two-Class (New Product Purchase) Discrimination Process by Using Multi-Criteria Linear Programming, in Kenneth D.Lawrence et al. (eds.). *Data mining Methods and Applications*. Auerbach Press of Taylor and Francis Publishing.

Lawrence, K., D. Pai, R. Klimberg, S. Lawrence. 2008. Understanding Donor Behavior: An Empirical Investigation by Using Parametric and Non-Parametric Classification Techniques, in Kenneth D.Lawrence et al. (eds.). *Business and Management Forecasting*. Jai Press, Elsevier Publishing, New York, NY.

Lawrence, K., D. Pai, R. Klimberg, S. Lawrence. 2009. Segmenting financial services market: An Empirical Study of Statistical and Non-parametric Methods, in Alice C.Lee and Cheng-Feu.Lee (eds.). *Handbook of Quantitative Finance*. Springer Verlag, New York, NY.

Lee, C., J. Ord. 1990. Discriminant analysis using least absolute deviations. *Decision Sciences*. 21 86-96.

Lee, K., D. Booth, P. Alam. 2005. A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*. 29(1) 1–16.

Li, R., Z. Wang. 2004. Mining classification rules using rough sets and neural networks. *European Journal of Operational Research*. 157(2) 439-448.

Lilliefors, H. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*. 62(318) 399-402.

Loucopoulos, C. 2001. Three-group classification with unequal misclassification costs-A mathematical programming approach. *Omega*. 29 291–297.

Loucopoulos, C., R Pavur. 1997. Computational characteristics of a new Mathematical programming model for the three-group Discriminant problem. *Computers and Operations Research*. 24(2) 179-191.

Loucopoulos, C., R. Pavur. 1997. Experimental evaluation of the classificatory performance of mathematical programming approaches to the three-group discriminant problem: The case of small samples. *Annals of Operations Research*. 74 191–209.

Maas, C., J. Hox. 2005. Sufficient Sample Sizes for Multilevel Modeling. *European Journal of Research Methods for the Behavioral and Social Sciences*. 1 85–91.

Mahmood, M., E. Lawrence. 1987. A performance analysis of parametric and nonparametric discriminant approaches to business decision making. *Decision Sciences*. 18 308–326.

Mangasarian, O. 1965. Linear and nonlinear separation of patterns by linear programming. *Operations Research*. 13 444–452.

Markowski, C., E. Markowski. 1987. An experimental comparison of several approaches to the discriminant problem with both qualitative and quantitative variables. *European Journal of Operational Research*. 28 74-78.

Markowski, C. 1990. On the balancing of error rates for LP discriminant methods. *Managerial and Decision Economics*. 11 235-241.

Meyers, L., G. Gamst, A. Guarino. 2006. Applied Multivariate Research: Design and Interpretation. *Sage Publications, Inc*. Thousand Oaks, CA.

Michie,D., J. Spiegelhalter, C. Taylor, eds. 1994. Machine Learning. *Neural and Statistical Classification*. Ellis Horwood series in Artificial Intelligence. Ellis Horwood, London.

Morgan, P., J. Teachman. 1988. Logistic Regression: Description, Examples, and Comparisons. *Journal of Marriage and the Family*. 50(4) 929-936.

Morrison, D. 1969. On the Interpretation of Discriminant Analysis. *Journal of Marketing Research*. 6 156-63.

Myers, J., E. Forgy. 1963. The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association*. 58(303) 799-806.

Nasir, M., R. John, S. Bennett. 2000. Predicting corporate bankruptcy using modular neural networks. *IEEE Computational Intelligence in Financial Engineering Conference*. New York, NY, 86-91.

Natter, M. 1999. Conditional market segmentation by neural networks: A Monte-Carlo study. *Journal of Retailing and Consumer Services*. 6 237-248.

Neter, J., W. Wasserman, M. Kutner. 1990. Applied Linear Statistical Models. 3rd edn. *Irwin*, Homewood, IL.

Nguyen, N., A. Cripps. 2001. Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*. 22(3) 313-336.

Odom, M., R. Sharda. 1990. A neural network model for bankruptcy prediction. *In Proceedings of the international joint conference on neural networks*. 2 163–168. IEEE Press, Alamitos, CA.

Ostermark, R., R. Hoglund. 1998. Addressing the multigroup discriminant problem using multivariate statistics and mathematical programming. *European Journal of Operational Research*. 108 224-237.

Paliwal, M., U. Kumar. 2009. Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*. 36 2–17.

Pavur, R. 1997. Dimensionality representation of linear discriminant function space for the multiple-group problem: An MIP approach. *Annals of Operations Research*. 74 37-50.

Pavur, R., C. Loucopoulos. 1995. Examining Optimal Criterion Weights in Mixed Integer Programming Approaches to the Multiple-Group Classification Problem. *The Journal of the Operational Research Society*. 46(5) 626-640.

Pavur, R., C. Loucopoulos. 2001. Evaluating the Effect of Gap Size in a Single Function Mathematical Programming Model for the Three-Group Classification Problem. *The Journal of the Operational Research Society*. 52(8) 896-904.

Pawlak, Z. 1991. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, New York, NY.

Pendharkar, P. 2002. A computational study on the performance of ANNs under changing structural design and data distributions. *European Journal of Operational Research.* 138 155–177.

Ragsdale, C., A. Stam. 1991. Mathematical Programming Formulations for the Discriminant Problem: An Old Dog Does New Tricks. *Decision Sciences*. 22 296–307.

Raudys, S., A. Jain. 1991. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 13(3) 252-264.

Ripley, B. 1994. Neural networks and related methods for classification. *Journal of the Royal Statistical Society (*Series B). 56(3) 409-456.

Robertson, T., J. Kennedy. 1968. Prediction of Consumer Innovators - Multiple Discriminant Analysis. *Journal of Marketing Research*. 5 64-69.

Rubin, P. 1990. A Comparison of Linear Programming and Parametric Approaches to the Two-Group Discriminant Problem. *Decision Sciences*. 21 373–386.

Sadat-Hashemi, S., A. Kazemnejad, C. Lucas, K. Badie. 2004. Predicting the type of pregnancy using artificial neural networks and multinomial logistic regression: a comparison study. *Neural Computing & Applications*. 14(3) 198-202.

Salchenberger, L., E. Cinar, N. Lash. 1992. Neural networks: A new tool for predicting thrift failures. *Decision Sciences*. 23 899–916.

Salzberg, S. 1991. A nearest hyperrectangle learning method. *Machine Learning*. 6 277–309.

Sharda, R. 1994. Neural networks for the MS/OR analyst: An application bibliography. *Interfaces*. 24(2) 116-130.

Shmueli, G., N. Patel, P. Bruce. 2006. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. *John Wiley and Sons, Inc*., NJ.

Short, R., K. Fukunaga. 1980. A New Nearest Neighbor Distance Measure. *Proceedings of the Fifth IEEE Computer Society Conference on Pattern Recognition and Image Processing*. IEEE Computer Society. Silver Spring, MD.

Simard, P., Y. LeCun, J. Denker, eds. 1993. Efficient pattern recognition using a new transformation distance. *Advances in Neural Information Processing Systems*. Morgan Kaufman. Los Altos, CA.

Sordo, M., Q. Zeng. 2005. On Sample Size and Classification Accuracy: A Performance Comparison. Springer, New York, NY.

Spear, N., M. Leis. 1997. Artificial neural networks and the accounting method choice in the oil and gas industry. *Accounting Management and Information Technology*. 7(3) 169–181.

Srinivasan, V., Y. Kim. (1987). Credit Granting: A Comparative Analysis of Classification Procedures. *The Journal of Finance*. 42(3) 665-681.

Stam, A., C. Ragsdale. 1992. On the classification gap in MP-based approaches to the discriminant problem. *Naval Research Logistics*. 39 545–559.

Sueyoshi, T., S. Hwang. 2004. A Use of Nonparametric Tests for DEA-Discriminant Analysis: A Methodological Comparison. *Asia-Pacific Journal of Operational Research* 21(2) 179-195.

Suh, E., K. Noh, C. Suh. 1999. Customer list segmentation using the combined response model. *Expert Systems with Application.* 17(2) 89–97.

Tam, K., M. Kiang. 1990. Predicting bank failures: A neural network approach. *Applied Artificial Intelligence*. 4(4) 265-282.

Tam, K., M. Kiang. 1992. Managerial applications of neural networks: The Case of Bank Failure Predictions. *Management Science*. 38(7) 926-947.

Tansey, R., M. White, R. Long. 1996. A comparison of loglinear modeling and logistic regression in management research. *Journal of Management*. 22(2) 339-358.

Thomas, L., K. Jung, S. Thomas, Y. Wu. 2006. Modeling consumer acceptance probabilities. *Expert Systems with Applications.* 30 499–506.

Utgoff, P. 1989. Perceptron trees: A case study in hybrid concept representations. *Connection Science*. 1 377-391.

Uysal, M., M. Roubi. 1999. Artificial neural networks versus multiple regression in tourism demand analysis. *Journal of Travel Research*. 38 111-118.

Walkling, R. 1985. Predicting Tender Offer Success: A Logistic Analysis. *Journal of Financial and Quantitative Analysis*. 20(4) 461-478.

Wedel, M., W. Kamakura. 1998. Market Segmentation: Conceptual and Methodological Foundations. Kluwer Academic Publishers, Inc. Boston, MA.

West, P., P. Brockett, L. Golden. 1997. A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*. 16(4) 370-391.

Westin, R. 1973. Predictions from binary choice models. Discussion paper no. 37. Northwestern University, Evanston, IL.

Wettschereck, D., T. Dietterich. 1995. An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*. 19 5–27.

White, H. 1989. An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks. *Proceedings of the International Joint Conference on Neural Networks*. 2 451-455.

Wiginton, J. 1980. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*. 15(3) 757-770.

Wilson, R., R. Sharda. 1994. Bankruptcy prediction using neural networks. *Decision Support Systems*. 11 545–557.

Wolpert, D. 1992. Stacked generalization, *Neural Networks*. 5 241– 259.

Zahavi, J., N. Levin. 1997a. Issues and problems in applying neural computing to target marketing. *Journal of Direct Marketing*. 11(4) 63-75.

Zahavi, J., N. Levin. 1997b. Applying neural computing to target marketing. *Journal of Direct Marketing*. 11(4) 76-93.

Zhang, G., B. Patuwo, M. Hu. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*. 14 35–62.

Zhang, G., M. Hu, B. Patuwo, D. Indro. 1999. Artificial neural networks in bankruptcy prediction - General framework and cross-validation analysis. *European Journal of Operational Research*. 116 16-32.

Zhang, Z., C. Zhang, eds. 2004. *Agent-based hybrid intelligent systems.* Springer-Verlag, Berlin, Germany.

Zopounidis, C., M. Doumpos. 2002. Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*. 138 229–246.

VITA

Dinesh Ramdas Pai

1971	Born December 3 in Mumbai, India.
1988	Secondary School Certificate (10^{th} Tenth Grade), St. Xavier's High School, Mumbai, India.
1990	Higher Secondary Certificate (12^{th} Grade), V.S.G. Technical Jr. College Mumbai, India.
1994	Bachelors degree in Mechanical Engineering, University of Pune, India.
1994-98	Project Engineer, Chemtex Engineering of India Limited, Mumbai, India.
1999	Summer Internship, Larsen & Toubro Limited, Mumbai, India.
2000	Master of Business Administration, University of Pune, India.
2000-01	Senior Marketing Executive, Reliance Industries Limited, Mumbai, India.
2001-02	Executive Assistant to the CEO, Bajaj Hindustan Limited, Mumbai, India.
20002-04	Independent Consultant, Mumbai, India.
2004-08	Teaching Assistant, Rutgers Business School, Rutgers University, Newark, New Jersey.
2008-09	Instructor, Rutgers Business School, Newark, New Jersey.
2009	Ph.D in Management, Rutgers Business School, Rutgers University, Newark, New Jersey.