Working Paper

Mathematical Programming Formulations for Two-group Classification with Binary Variables

*Ognian K. Asparoukhov**
*Antonie Stam***

WP-96-92
August 1996

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Mathematical Programming Formulations for Two-group Classification with Binary Variables

*Ognian K. Asparoukhov**
*Antonie Stam***

WP-96-92
August 1996

*Centre of Biomedical Engineering, Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl. 105, 1113 Sofia, Bulgaria

**Department of Management, Terry College of Business, The University of Georgia, Athens, GA 30602, U.S.A.
and
International Institute for Applied Systems Analysis
Laxenburg, Austria

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

MATHEMATICAL PROGRAMMING FORMULATIONS FOR TWO-GROUP CLASSIFICATION WITH BINARY VARIABLES

ABSTRACT

In this paper, we introduce a nonparametric mathematical programming (MP) approach for solving the binary variable classification problem. In practice, there exists a substantial interest in the binary variable classification problem. For instance, medical diagnoses are often based on the presence or absence of relevant symptoms, and binary variable classification has long been used as a means to predict (diagnose) the nature of the medical condition of patients. Our research is motivated by the fact that none of the existing statistical methods for binary variable classification – parametric and nonparametric alike – are fully satisfactory.

The general class of MP classification methods facilitates a geometric interpretation, and MP-based classification rules have intuitive appeal because of their potentially robust properties. These intuitive arguments appear to have merit, and a number of research studies have confirmed that MP methods can indeed yield effective classification rules under certain non-normal data conditions, for instance if the data set is outlier-contaminated or highly skewed. However, the MP-based aproach in general lacks a probabilistic foundation, an *ad hoc* assessment of its classification performance.

Our proposed nonparametric mixed integer programming (MIP) formulation for the binary variable classification problem not only has a geometric interpretation, but also is consistent with the Bayes decision theoretic approach. Therefore, our proposed formulation possesses a strong probabilistic foundation. We also introduce a linear programming (LP) formulation which parallels the concepts underlying the MIP formulation, but does not possess the decision theoretic justification.

An additional advantage of both our LP and MIP formulations is that, due to the fact that the attribute variables are binary, the training sample observations can be partitioned into multinomial cells, allowing for a substantial reduction in the number of binary and deviational variables, so that our formulation can be used to analyze training samples of almost any size.

We illustrate our formulations using an example problem, and use three real data sets to compare its classification performance with a variety of parametric and nonparametric statistical methods. For each of these data sets, our proposed formulation yields the mimimum possible number of misclassifications, both using the resubstitution and the leave-one-out method.

Keywords: Binary Variables, Classification Analysis, Discriminant Analysis, Linear Programming, Mixed Integer Programming.

Acknowledgements: The first author was supported partially by the National Science Fund of Bulgaria, grant reference number 1854. The second author gratefully acknowledges the support provided by the Terry College of Business of the University of Georgia through a Terry Summer Research Grant.

MATHEMATICAL PROGRAMMING FORMULATIONS FOR TWO-GROUP CLASSIFICATION WITH BINARY VARIABLES

1. INTRODUCTION

Over the years, a considerable body of literature has accumulated on classification analysis, with its usefulness demonstrated in various fields, including engineering, medical and social sciences, economics, marketing, finance and management (Anderson *et al.* 1972; McLachlan 1992; Joachimsthaler and Stam 1988, 1990; Ragsdale and Stam 1992; Huberty 1994; Yarnold *et al.* 1994). Most of the research in classification analysis is based on statistical methods (Dillon and Goldstein 1978; Hand 1981; McLachlan 1992; Huberty 1994). However, the classification performance of existing parametric and nonparametric statistical methods has not been fully satisfactory. For instance, it is well-documented that parametric statistical methods, such as Fisher's linear discriminant function (LDF) (Fisher 1936) and Smith's quadratic discriminant function (QDF) (Smith 1947) may yield poor classification results if the assumption of multivariate normally distributed attributes is violated to a significant extent (McLachlan 1992; Huberty 1994; Krzanowski 1988; Joachimsthaler and Stam 1990; Duarte Silva 1995). As we will discuss in more detail below, nonparametric methods may give overly positive resubstitution (training sample classification) rates, while performing poorly on validation samples, and may be overly sensitive to certain data conditions (Goldstein and Dillon 1978; Hand 1983, 1993; McLachlan 1992).

A number of the statistical classification methods are based on distance measures. Some involve probability density functions and variance-covariances, and have a Bayes decision theoretic probabilistic interpretation, while others have a geometric interpretation only. An example of a distance-based measure is the Euclidean distance measure, which obviously has a goemetric interpretation. If the attribute variables are independent, the Euclidean distance measure is equivalent to the Mahalanobis distance, with the usual probabilistic interpretation. However, if the variables are correlated the Euclidean measure does not have a probabilistic justification, as it does not involve any function of the probability density functions.

Recently, a class of nonparametric mathematical programming (MP)-based techniques has attracted considerable research attention. Among the most widely known MP methods are the minimize the sum of deviations (MSD) method (Freed and Glover 1981b; Mangasarian 1965; Minnick 1961; Smith 1968), the minimize the maximum deviation (MMD) method (Freed and Glover 1981a; Rubin 1989), the minimize the number of misclassifications (MIP) method (Ibaraki and Muroga 1970; Liitschwager and Wang 1978; Asparoukhov 1985), and the Hybrid method (Glover, Keene and Duea 1988). Nonlinear (Stam and Joachimsthaler 1989), multi-group (Gehrlein 1986; Gochet *et al.* 1996), polynomial and second-order variants (Banks and Abad 1994; Rubin 1994; Duarte Silva and Stam 1994; Wanarat and Pavur 1996) of linear MP formulations have been proposed as well. MP-based methods for classification have a geometric interpretation, and are based on the construction of surfaces that provide an optimal separation of the groups. The optimization criterion either minimizes some function of the undesirable distances of the training sample observations from the separating surface, or

minimizes the number of misclassified observations directly. Most MP-based methods lack a probabilistic justification. An exception is the method proposed by Lam *et al.* (1993).

In this paper, we focus on two-group classification problems with binary attribute variables. There are numerous real-life binary variable classification problems, *e.g.*, in the field of medical disease diagnosis, where the medical condition of patients is evaluated on the basis of the presence or absence of relevant symptoms. Some examples of such applications will be analyzed and discussed in Section 5. It is obvious that the multivariate distribution of the binary attributes is non-normal, and it appears promising to analyze such problems using nonparametric approaches (like MP ones). A number of specialized statistical discriminant methods have been developed for problems with categorical (usually binary) or mixed (continuous and binary) variables (Goldstein and Dillon 1978; Hand 1981; Krzanowski 1993; McLachlan 1992; Huberty 1994). To date, there has been little MP-based research in this area, with the exception of Markowski and Markowski (1987), who discussed the mixed variable problem, and Stam and Joachimsthaler (1990) and Stam and Jones (1990), who included discrete uniform data conditions in their simulation experiments.

Our purpose is to develop an MP-based formulation for the binary variable classification problem which is fully consistent with the Bayes decision theoretic approach, and has both intuitive appeal and a formal probabilistic justification. In Section 2, we define the classification problem formally, and present the decision theoretic approach to classification. Section 3 reviews existing statistical methods for binary variable classification. In Section 4, we derive a simple Bayes decision theoretic classification rule for the case of binary attribute variables, and use this rule to formulate the Bayes decision theoretic MIP method (BMIP). Section 5 presents a rigorous analysis of three real data sets, comparing the classificatory performance of the BMIP and a related MSD method with a plethora of existing parametric and nonparametric statistical methods. Section 6 contains concluding comments.

2. DECISION THEORETIC APPROACH

Consider the classification problem with r mutually exclusive and collectively exhaustive groups, and denote group j by G_j. Suppose that the characteristics of each group are described by the p-dimensional attribute vector $\mathbf{x} = (x_1, ..., x_p)^T$. The purpose in classification analysis is to predict the group membership of an observation i based on the characteristics of its vector of attribute values $\mathbf{x}_i$. Define the conditional probability that i belongs to G_j by $p_j(\mathbf{x}_i) = p(\mathbf{x}_i \mid G_j)$, the prior probability of membership in G_j by $p(G_j)$, the posterior probability of group membership by $p(G_j \mid \mathbf{x}_i)$, and the cost associated with erroneously classifying an observation from G_l into G_j by c_{jl}. Most decision theoretic classification rules are based on the posterior probabilities. These probabilities are usually unknown, but may be estimated using Bayes' theorem through $\hat{p}(G_j \mid \mathbf{x}_i) = \hat{p}(G_j)\hat{p}_j(\mathbf{x}_i)/\hat{p}(\mathbf{x}_i)$, where $\hat{p}(\mathbf{x}_i)$ is calculated as $\hat{p}(\mathbf{x}_i) = \sum_j \hat{p}(G_j)\hat{p}_j(\mathbf{x}_i)$.

The Bayes decision theoretic approach to classification seeks to divide the attribute space $X \subset \Re^p$ into r mutually exclusive, collectively exhaustive regions R_1, ..., R_r, such that observation i

will be assigned to G_j if and only if $\mathbf{x}_i \in R_j$ (Anderson 1984; Das Gupta 1973). Among the most widely known decision theoretic rules are the maximum likelihood rule, the rule which minimizes the expected misclassification costs, and the rule which minimizes the total probability of misclassification. The Bayes rule in (2.1) minimizes the expected cost of misclassification (Hand 1981),

$$\text{Classify observation } i \text{ into } G_k \text{ iff } \sum_{j=1, j\neq k}^{r} c_{kj}p(G_j)p_j(\mathbf{x}_i)/p(\mathbf{x}_i) < \sum_{j=1, j\neq q}^{r} c_{qj}p(G_j)p_j(\mathbf{x}_i)/p(\mathbf{x}_i), \quad q = 1, \ldots, r;\ q \neq k. \tag{2.1}$$

As $p(\mathbf{x}_i)$ is common to both sides, it can be omitted from (2.1). If the misclassification costs across groups are equal, the Bayes discriminant rule minimizing the total probability of misclassification (Hand 1981) classifies observation i into G_k for which (2.2) holds,

$$\text{Classify observation } i \text{ into } G_k \text{ iff } p(G_k)p_k(\mathbf{x}_i)/p(\mathbf{x}_i) = \max_j\{p(G_j)p_j(\mathbf{x}_i)/p(\mathbf{x}_i)\}. \tag{2.2}$$

The rule in (2.2) in fact classifies an observation into the group with the highest posterior probability. As in (2.1), the term $p(\mathbf{x}_i)$ is usually omitted from the expression. The maximum likelihood discriminant rule in (2.3) assigns observation i to G_k with the highest probability density function value,

$$\text{Classify observation } i \text{ into } G_k \text{ iff } p_k(\mathbf{x}_i) = \max_j\{p_j(\mathbf{x}_i)\}. \tag{2.3}$$

In the remainder of this paper, we will focus on building decision theoretic rules for the binary variable classification problem based on the general decision rule in (2.2). Specifically, we will develop MP-based rules which are consistent with this decision-theoretic rule. Although we will limit ourselves to rules of the type of (2.2), it is straightforward to develop analogous MP-based rules based on (2.1) and (2.3).

3. CLASSIFICATION IN THE PRESENCE OF BINARY VARIABLES: STATISTICAL METHODS

Consider the case where observation i is characterized by the binary attribute vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, *i.e.*, $x_{iu} = 0$ or 1, for all u. Suppose that the number of training sample observations in G_j is given by n_j, for a total of $n = \sum_{j=1}^{r} n_j$ observations in the sample. Thus, a training sample observation i from G_j is characterized by the p-tuple $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$ of binary variables.

Define the degree of discordance between two binary vectors $\mathbf{x}$ and $\mathbf{y}$ by the Hamming distance measure $d(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y})$, and let the number of training sample observations $\mathbf{x}_{ij}$ with $d(\mathbf{x}, \mathbf{x}_{ij}) = k$ be given by $n_{jk}(\mathbf{x})$. In the case of binary vectors, $d(\mathbf{x}, \mathbf{y})$ represents the number of components of $\mathbf{x}$ and $\mathbf{y}$ which differ in value (*i.e.*, 0 vs. 1 or 1 vs. 0). Denote the number of observations $\mathbf{x}_{ij}$ located in the multinomial cell of a given $\mathbf{x}$, *i.e.*, with $d(\mathbf{x}, \mathbf{x}_{ij}) = 0$, by $n_{jo}(\mathbf{x})$. Since $\mathbf{x}$ consists of binary components, we can associate each distinct $\mathbf{x}^{\mathrm{T}} = (x_1, \ldots, x_p)$ with uniquely with cell $\mathcal{C}(s) = 1 + \sum_{i=1}^{p} x_i 2^{(i-1)}$, $s = 1, \ldots, t$, where $t = 2^p$. For example, all observations in cell $\mathcal{C}(1)$ have

$x_1 = ... = x_p = 0$, whereas all observations in cell $\mathcal{C}(2)$ have $x_1 = 1$, $x_2 = ... = x_p = 0$. Denote the attribute vector associated with each observation located in a given cell $\mathcal{C}(s)$ by $\mathbf{b}_s$. Obviously, any classification rule based on binary attribute variables will allocate *all* observations in cell $\mathcal{C}(s)$ to the *same* group. Denote the number of training sample observations $i \in G_j$ which are located in cell $\mathcal{C}(s)$ by n_{js}. Then, in the case of two groups the number of misclassified observations for cell s will be either n_{1s} (if the decision rule assigns the observations in cell $\mathcal{C}(s)$ to G_2) or n_{2s} (if the decision rule is to assign the observations in cell $\mathcal{C}(s)$ to G_1).

Next, we review a number of widely used statistical methods for binary variable classification. Each of these methods estimates the group-conditional distribution $p_j(\mathbf{x}_i) = p_j(\mathbf{x}_i \mid G_j)$, $j = 1, ..., r$, and classifies an observation i into the group G_k with the highest estimated posterior probability $\hat{p}(G_k \mid \mathbf{x}_i)$, determined from the $\hat{p}_j(\mathbf{x}_i)$ using Bayes' theorem.

Full Multinomial, Nearest Neighbor and Kernel Methods

The *full multinomial procedure* estimates the $p_j(\mathbf{x})$ by the relative frequencies $\hat{p}_j(\mathbf{x}) = n_{jo}(\mathbf{x})/n_j$, $j = 1, ..., r$. For a meaningful statistical interpretation, it is recommended that each cell in the experimental design contain at least five training sample observations. In practice, frequently some of the cells are empty, in particular if the number of attributes is large or if the number of observations is small. The full multinomial procedure requires many parameters and provides little information about the shape of the distribution $p_j(\mathbf{x})$. On the positive side, the procedure is straightforward, easy to understand and easy to implement, and yields an asymptotically optimal, unique minimum variance unbiased estimator of $p_j(\mathbf{x})$ (Hand 1993).

Hills (1967) proposes a smoothed group membership probability estimator for the binary variable classification problem, $\hat{p}_j(\mathbf{x}) = n_j^{-1} \sum_{h=0}^{L} n_{jh}(\mathbf{x})$, $j = 1, ..., r$, where L, $0 \leq L \leq p$, is the order of the procedure. This estimator, known as Hill's k-nearest neighbor estimator (*kNN-Hills*), avoids the problem with empty cells by including in the numerator the training sample observations in all cells with a Hamming distance from $\mathbf{x}$ of at most L. Note that the full multinomial is the kNN-Hills estimator for L =0. While intuitively appealing, the kNN-Hills estimator is *ad hoc* and lacks a theoretical or model-based justification (Krzanowski 1988).

The *kNN-Hall* estimator (Hall 1981b) is an adaptive variant of the kNN-Hills estimator of the form $\hat{p}_j(\mathbf{x}) = n_j^{-1} \sum_{h=0}^{L} w_{jh} n_{jh}(\mathbf{x})$. The weights w_{jh} are chosen to minimize $\Delta(w_{j1}, ..., w_{jL})^{\mathrm{T}} = \sum_{s=1}^{t} E\left\{\hat{p}_j(\mathbf{b}_s) - p_j(\mathbf{b}_s)\right\}^2$. Unfortunately, the kNN-Hall estimator may yield negative probability estimates, but this usually arises only if the true probabilities are small or if the training sample is too small. Negative estimates can be interpreted as a warning that the probabilities in question cannot be estimated accurately given the current design (Hall 1981b).

Aitchison and Aitkin (1976) propose the nonparametric *kernel estimator*, $\hat{p}_j(\mathbf{x}; \lambda) = n_j^{-1} \sum_{h=1}^{n_j} \lambda^{p - d(\boldsymbol{x}_{hj}, \boldsymbol{x})} (1-\lambda)^{d(\boldsymbol{x}_{hj}, \boldsymbol{x})}$, where λ is a smoothing parameter such that $0.5 \leq \lambda \leq 1$. These authors suggest to estimate λ by cross-validation, for instance by means of the leave-one-out (LOO) method, using an estimate of the likelihood function. Unfortunately, the resulting adaptive estimator

may behave erraticly in the presence of empty or near-empty cells. To overcome this difficulty, Hall (1981a) proposes an alternative estimator which minimizes a global function of the mean squared error.

Kernel and kNN estimators are based on one single assumption, namely that neighboring cells are highly correlated, so that adjacent cells will have similar group membership probabilities. In contrast, no such assumption is made in the full multinomial method (Hand 1981). The kNN-Hall and kernel estimators are very flexible and have a tendency to overfit the data (Aitchison and Aitkin 1976; Asparoukhov and Andreev 1995; Hall 1981b). As a consequence, these methods tend to yield overly optimistic and heavily biased resubstitution (*i.e.*, training sample) misclassification errors, thus providing an unreliable measure of classification accuracy on validation samples. Moreover, these methods have been found to be effective only if the training sample is large or if the number of variables is large (Asparoukhov and Andreev 1995; Hand 1983).

Bahadur Model

The first-order Bahadur model (Bahadur 1961) assumes that the variables are independent, and estimates $p_j(\mathbf{x}_i)$ by $\widehat{p}_j(\mathbf{x}_i) = \prod_{i=1}^{p} (\theta_{ij})^{x_i}(1-\theta_{ij})^{1-x_i}$, where $\theta_{ij} = p(x_i = 1 \mid G_j)$. This model involves few parameters and can easily handle missing data, but tends to be overly optimistic (*i.e.*, biased) in terms of estimating the group membership probabilities (Hand 1993), and the classification accuracy of the first-order Bahadur model may decrease significantly if the variables are correlated. The second-order Bahadur model, which uses first order correlation terms, performs much better if the variables are correlated (Dillon and Goldstein 1978).

Log Linear Models (LLM), Logistic and Quadratic Regression, and Normality-Based Procedures

Log linear models (*LLM*) are widely used for analyzing contingency tables, and estimate $\log(p_j(\mathbf{x}))$ as a linear function of the main effects and interactions between the binary variables (Agresti 1990). However, the decision of which main effects and interactions to include in the analysis has to be made prior to fitting the model, and empty cells may cause serious estimation problems. As a result, this method is of limited use for analyzing classification problems with binary variables, particularly if the number of variables (and therefore the number of cells) is large.

The logistic regression (*LR*) and quadratic logistic regression (*QLR*) methods avoid the problem of estimating the density function, by assuming that $p(G_j \mid \mathbf{x})$ has a logistic form (Cox 1966; Day and Kerridge 1967; Anderson 1972, 1975). In the LR method, the interaction structure between the different groups is assumed to be equal, whereas the QLR assumes a more general structure, albeit at the expense of having to use higher-dimensional iterative estimation schemes (Anderson 1975).

If the observations are multivariate normally distributed with equal variance-covariances across groups, Fisher's (1936) linear discriminant function (*LDF*) yields the optimal classification rule (Anderson 1984). Similarly, if the observations are multivariate normally distributed with unequal variance-covariances across groups, Smith's (1947) quadratic discriminant function (*QDF*) is optimal (Anderson 1984). Moore (1973) demonstrates that the LDF tends to perform better than the QDF.

This may be due to the large number of parameters to be estimated in the QDF, which plays an important role if the training sample is small relative to the number of attributes (Duarte Silva 1995).

As the LDF is relatively robust and easy to apply, it has often been used to analyze classification problems for which the normality assumption is mildly violated (Krzanowski 1977), for instance in the case of binary attribute variables (Dillon and Goldstein 1978; Gilbert 1968; Hand 1983; Moore 1973; Krzanowski 1977; Trampisch 1978). One characteristic of the LDF is the stability of its classification performance as the number of variables increases. Dillon and Goldstein (1978) recommend the use of the LDF in situations of moderate correlations and reasonably large mean differences. A number of authors have studied the performance of LR in relation to the LDF (McLachlan 1992). The general consensus is that logistic discrimination is preferred to the LDF if the distributions are clearly non-normal (as with binary variables) or the dispersion matrices are strongly unequal (Krzanowski 1988).

Fourier Procedure

Ott and Kronmal (1976) propose a nonparametric model based on an orthogonal expansion of the density in terms of discrete *Fourier* series. This model, however, is unsuitable for problems with a large number of attribute variables (Asparoukhov and Andreev 1995; Titterington *et al.* 1981).

All of the statistical methods described above are included in the experimental comparison below. Other statistical methods for binary classification that we will not discuss in detail, as these are not part of our study, include single-stage methods such as the minimum logit χ^2, minimax estimators, Rademacher-Walsh polynomial approximations (Goldstein and Dillon 1978; Hand 1981, 1982, 1983; Martin and Bradly 1972; McLachlan 1992), neural networks (Lippmann 1989; Másson and Wang 1990; Rypley 1994; Salchenberger *et al.* 1992; Tam and Kiang 1992), classification trees (Hartigan 1975), and multi-stage methods such as multiclassifiers (Xu *et al.* 1992), multilevel classifiers (Benediktson and Swain 1992; Ng and Abramson 1992) and tree classifiers (Sturt 1980).

4. MP-BASED CLASSIFICATION FOR THE BINARY VARIABLE PROBLEM

If the attribute variables are binary, the r groups may be thought of as swarms of points in $\Re^p$. An observation signifies one single point in $\Re^p$, and it is intuitively attractive to allocate it to the "closest" training sample group (Krzanowski 1988). The Bayes decision theoretic approach to classification seeks to divide the attribute space into regions that minimize either the expected misclassification cost or the total misclassification probability, or maximize the likelihoods, based on probability density functions. Hand (1981) notes that, although perhaps intuitively attractive, non-Bayesian decision rules may result in poor classifiers, unless the decision surface closely resembles that of the Bayes rule.

With the exception of the MIP, which directly minimizes either the number of misclassified training sample observations or the expected misclassification costs, MP methods use distance measures from the boundaries of R_j for classification purposes. Most MP methods are based on the absolute distance criterion (L_1 norm distances), which derives its intuitive appeal as a potentially robust alternative to L_2 norm-based parametric normality-based classification methods such as the LDF and QDF if the data are clearly non-normal. A number of simulation studies have shown that for non-normal and outlier-contaminated classification problems, several MP-based methods, in particular the MSD and Hybrid methods, may indeed give better classification results than the LDF, QDF, LR, QLR, kernel and Nearest Neighbor methods (Glorfeld and Kattan 1989; Joachimsthaler and Stam 1990; Duarte Silva and Stam 1994; Duarte Silva 1995), but not all research studies confirm this finding. In the case of binary variable classification problems, the normality assumption is clearly violated, so that MP methods appear natural candidates for solving these problems.

4.1. Developing a Bayes Decision Theoretic Rule for the Binary Variable Discriminant Problem

We next develop the MIP-based Bayesian rule (BMIP) is optimal not only in a geometric sense, but also in the Bayes decision theoretic sense. Although we could develop analogous MP-based classification rules based on (2.1) and (2.3), in this paper we will focus on the Bayes discriminant rule in (2.2) which minimizes the total probability of misclassification. In the case of binary vectors $\mathbf{x}_i$, $p_j(\mathbf{x}_i) = p(\mathbf{x}_i \in \mathcal{C}(s) \mid G_j)$, and for the two-group problem (2.2) can be written as (4.1),

$$\begin{aligned}\text{Classify observation } i \text{ into } G_1, \text{ if } p(G_1)p(\mathbf{x}_i \in \mathcal{C}(s) \mid G_1) \geq p(G_2)p(\mathbf{x}_i \in \mathcal{C}(s) \mid G_2),\\ \text{and into } G_2 \text{ otherwise,}\end{aligned} \tag{4.1}$$

The rule in (4.1) in fact maximizes the posterior probability of group membership $p(G_j \mid \mathbf{x}_i) = p(G_j)p_j(\mathbf{x}_i)/p(\mathbf{x}_i)$, but we can omit $p(\mathbf{x}_i) > 0$ from the expression because it is common to both the left- and right-hand-side of the inequality. The rule in (4.1) shows that, in order to classify observation i, we only need to identify the group G_j for which the posterior probability is the highest, regardless of how much higher it is, or what the exact probability values are. In other words, in classifying observation i, it makes no difference whether $p(G_1 \mid \mathbf{x}_i \in \mathcal{C}(s)) = 0.99$ and $p(G_2 \mid \mathbf{x}_i \in \mathcal{C}(s)) = 0.01$, or $p(G_1 \mid \mathbf{x}_i \in \mathcal{C}(s)) = 0.51$ and $p(G_2 \mid \mathbf{x}_i \in \mathcal{C}(s)) = 0.49$; in both cases, observation i is assigned to G_1.

We can estimate $p_j(\mathbf{x}_i)$ by the relative frequencies in the training sample, yielding the unbiased estimator in (4.2),

$$\hat{p}_j(\mathbf{x}_i) = \hat{p}(\mathbf{x}_i \in \mathcal{C}(s) \mid G_j) = n_{js}/n_j. \tag{4.2}$$

If the prior group membership probabilities $p(G_j) = q_j$ are known, the joint probabilities are estimated by $\hat{\mathrm{p}}(\mathbf{x}_i \in \mathcal{C}(s) \cap \mathbf{x}_i \in G_j) = q_j n_{js}/n_j$, and (4.1) can be written as (4.3),

$$\text{Classify observation } i \in \mathcal{C}(s) \text{ into } G_1, \text{ if } \frac{q_1 n_{1s}}{n_1} \geq \frac{q_2 n_{2s}}{n_2}, \text{ and into } G_2 \text{ otherwise.} \tag{4.3}$$

If unknown in advance, the prior probabilities can be estimated by the relative frequencies of training sample observations in G_j, yielding $\hat{p}(G_j) = n_j/(n_1 + n_2)$, $j = 1, 2$, so that the decision rule is to assign observation $i \in \mathcal{C}(s)$ to G_1, if $\frac{n_{1s}}{n_1 + n_2} \geq \frac{n_{2s}}{n_1 + n_2}$, and to G_2 otherwise. Canceling common terms, the Bayesian decision theoretic rule can be simplified to (4.4),

$$\text{Classify observation } i \in \mathcal{C}(s) \text{ into } G_1, \text{ if } n_{1s} \geq n_{2s}, \text{ and into } G_2 \text{ otherwise.} \tag{4.4}$$

The rule in (4.4) indicates that we can use the number of observations in the different cells to estimate the posterior group membership probabilities, and therefore to predict the group membership of each observation. Since (4.4) was drived directly from the Bayes decision theoretic classification rule in (2.2), this approach is not just *ad hoc* but has a strong decision theoretical justification.

Inequalities of the type of (4.4) are easily implemented within the MP context. Hence, within the Bayes decision theoretic framework we should focus our attention on the fitting of inequalities, rather than on the estimation of the probability density function. We will refer to this approach as the BFI (Bayesian fitting of inequalities). Rather than maximizing the posterior probability *directly*, the BFI approach seeks to maintain the correct direction of the inequalities according to (4.4) for as many training sample observations as possible, therefore implicitly maximizing the posterior probability of group membership.

In the MP context, (4.4) implies that we have n inequalities, one for each training sample observation, and fit the direction of each inequality as a function of the binary attribute variables. Denoting the classification function by $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$, where $\mathbf{w}$ is the p-dimensional vector of attribute coefficients that are to be estimated, the classification score of observation i by $f(\mathbf{x}_i)$, and the cut-off value separating G_1 and G_2 by c, the MP formulation will classify i into G_1 if $f(\mathbf{x}_i) \leq c$, and into G_2 if $f(\mathbf{x}_i) > c$. The classification function $f(\mathbf{x})$ may either be linear in the attributes x_i or nonlinear functions of the attributes (*e.g.*, quadratic or polynomial). In order to keep the notation simple, we will limit our notation to the original attribute vector $\mathbf{x}$. Of course, in either case the resulting formulation is fully consistent with the BFI approach.

4.2. MIP Formulations for the Binary Variable Discriminant Problem

We are now ready to formulate a MIP formulation for the general two-group binary variable classification problem based on the BFI approach in (4.4), and derive an equivalent but greatly simplified formulation with many less binary variables and constraints, taking advantage of the special structure of the binary variable classification problem.

The general MIP formulation is given as Problem I,

Problem I:

$$\text{Minimize } z_1 = \sum_{i=1}^{n} \delta_i$$

Subject to:

$$\mathbf{x}_i^{\mathrm{T}}\mathbf{w} - M\delta_i \leq c, \qquad i \in G_1,$$

$$\mathbf{x}_i^{\mathrm{T}}\mathbf{w} + M\delta_i > c, \qquad i \in G_2,$$

$$w_k \text{ and } c \text{ are unrestricted, } k = 1, \ldots, p,$$

$$\delta_i = 1 \text{ if observation } i \text{ is misclassified, and } \delta_i = 0, \text{ otherwise, } i = 1, \ldots, n,$$

where M is a sufficiently large positive scalar.

It is well-documented that this formulation, with $n = n_1 + n_2$ proper constraints and n binary variables, is computationally feasible for small training samples only (Stam and Joachimsthaler 1990; Koehler and Erenguc 1990; Banks and Abad 1991; Soltysik and Yarnold 1993, 1994; Duarte Silva 1995). However, as in our case all observations located in a given cell $\mathcal{C}(s)$ have identical values for each of the p binary variables and will be classified into the same group, we can combine the problem constraints and contributions to the objective function z_1 for each cell. Replacing the individual observations $\mathbf{x}_i \in \mathcal{C}(s)$ by the corresponding vector $\mathbf{b}_s$ (recall that $\mathbf{b}_s = \mathbf{x}_i$ iff $i \in \mathcal{C}(s)$), Problem I can be restated as:

Problem II:

$$\text{Minimize } z_2 = \sum_{s=1,\ \mathcal{C}(s) \neq \emptyset}^{t} \left\{ n_{1s}\delta_{1s} + n_{2s}\delta_{2s} \right\}$$

Subject to:

$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w} - M\delta_{1s} \leq c, \quad \text{if } n_{1s} > 0,\ s = 1, \ldots, t,$$

$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w} + M\delta_{2s} > c, \quad \text{if } n_{2s} > 0,\ s = 1, \ldots, t,$$

$$w_k \text{ and } c \text{ are unrestricted, } k = 1, \ldots, p,$$

$$\delta_{js} = 1 \text{ if the observations from } G_j \text{ in cell } \mathcal{C}(s) \text{ are misclassified,}$$

$$\text{and } \delta_{js} = 0, \text{ otherwise, } s = 1, \ldots, t;\ j = 1, 2.$$

Problem II has at most two binary variables and at most two proper constraints for each cell $\mathcal{C}(s)$, for a total of at most $2t$ binary variables and at most $2t$ proper constraints. The optimal solutions to Problems I and II are identical. As we need to include a constraint only if the corresponding $n_{js} > 0$, the actual number of binary variables and constraints may be strictly less than $2t$.

Problem II can be tightened further, because $\mathbf{b}_s^{\mathrm{T}}\mathbf{w} - M\delta_{js}$ will either be at most c or exceed c, and either all training sample observations $i \in \mathcal{C}(s)$ that belong to G_1 will be classified correctly, *i.e.*, $\delta_{1s} = 0$ and $\delta_{2s} = 1$, or those belonging to G_2 will be classified correctly, in which case $\delta_{1s} = 1$ and $\delta_{2s} = 0$. Note that $\delta_{1s}\delta_{2s} = 0$ and $\delta_{1s} + \delta_{2s} = 1$, for each s. Using the Bayes decision theoretic rule in (4.4), we can minimize the the total probability of misclassification by assigning all observations in $\mathcal{C}(s)$ to G_1 if $n_{1s} \geq n_{2s}$, and to G_2 otherwise. Therefore, we need only one constraint for each cell, rather than the two constraints used in Problem II.

If $n_{1s} \geq n_{2s}$, the component of z_2 associated with $\mathcal{C}(s)$ becomes $n_{1s}\delta_{1s} + n_{2s}\delta_{2s} = (n_{1s} - n_{2s})\delta_{1s} + n_{2s}$. Similarly, if $n_{1s} < n_{2s}$ this component of z_2 equals $n_{1s}\delta_{1s} + n_{2s}\delta_{2s} =$

$(n_{2s}-n_{1s})\delta_{2s}+n_{1s}$. Therefore, the contribution of $\mathcal{C}(s)$ to z_2 equals $|\, n_{1s}-n_{2s}\,|\,\delta_s+\min(n_{1s},\ n_{2s})$, where the binary variable δ_s equals 1 iff the *majority* of training sample observations in $\mathcal{C}(s)$ is misclassified. Hence, the objective function component for $\mathcal{C}(s)$ is weighted according to the difference between the number of observations in $\mathcal{C}(s)$ that belong to each group. For each individual cell $\mathcal{C}(s)$, the minimum number of misclassified observations equals $\min(n_{1s},\ n_{2s})$.

Based on the above, we rewrite Problem II as the BFI formulation in Problem III:

Problem III: (BMIP)

$$\text{minimize } z_3 = \sum_{s=1,\mathcal{C}(s)\neq\emptyset}^{t} \Big\{ |\, n_{1s}-n_{2s}\,|\,\delta_s+\min(n_{1s},\ n_{2s})\Big\}, \tag{4.5}$$

Subject to:

$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w}-M\delta_s \le c, \qquad \text{if } n_{1s}\ge n_{2s};\ n_{1s}>0,\ s=1,\ ...,\ t,$$

$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w}+M\delta_s > c, \qquad \text{if } n_{1s}< n_{2s},\ s=1,\ ...,\ t,$$

$$w_k \text{ and } c \text{ are unrestricted, } k=1,\ ...,\ p,$$

$\delta_s = 1$ if the majority of observations $i\in\mathcal{C}(s)$ is misclassified, and $\delta_s=0$, otherwise, $s=1,\ ...,\ t$.

Problem III has at most t binary variables and at most t proper constraints, and has the same optimal solution as Problem II. Note that, if $\delta_s=0$ for all s, $z_3^*=\sum_{s,\mathcal{C}(s)\neq\emptyset}\min(n_{1s},\ n_{2s})$ equals the minimum number of misclassifications, which term is a constant and can be omitted from the objective function. Also note that, since the value of each n_{js} in the training sample is known *a priori* at the time of the model formulation, the objective function coefficients are determined prior to the analysis.

Table 1 Here

As an example, consider the two-group classification problem in Table 1 with $p=2$ binary attributes, for a total of $t=2^p=4$ cells. From Table 1, we see that $n_1=50$, $n_2=40$, and the training sample size equals $n=90$. The third and fourth columns show the distribution of the training sample observations over the different cells. For instance, of the 21 observations located in cell $\mathcal{C}(1)$ 15 belong to G_1 ($n_{11}=15$) and 6 to G_2 ($n_{12}=8$), so that $|\, n_{11}-n_{21}\,| = 15-6=9$ and $\min(n_{11},\ n_{21})=6$. The BMIP fomulation according to Problem III is as follows,

$$\begin{gathered}
\text{Minimize } z_3 = 9\delta_1+9\delta_2+6\delta_3+16\delta_4+25\\
0\times w_1+0\times w_2-M\delta_1 \le c, \qquad \text{for } \mathcal{C}(1),\\
0\times w_1+1\times w_2+M\delta_2 > c, \qquad \text{for } \mathcal{C}(2),\\
1\times w_1+0\times w_2+M\delta_3 > c, \qquad \text{for } \mathcal{C}(3),\\
1\times w_1+1\times w_2-M\delta_4 \le c, \qquad \text{for } \mathcal{C}(4),\\
w_1,\ w_2 \text{ and } c \text{ are unrestricted,}\\
\delta_s=1, \text{ if the majority of observations in } \mathcal{C}(s) \text{ is misclassified,}\\
\text{and } \delta_s=0, \text{ otherwise, } s=1,\ ...,\ 4.
\end{gathered}$$

The greatly simplified formulation in Problem III renders the BMIP approach computationally feasible for *any* size training sample.

4.3. MSD Formulations for the Linear Binary Variable Discriminant Problem
Development of the Conventional MSD Formulation

In this section, we develop a reduced size MSD formulation for the binary variable classification problem, much analogous to the BMIP formulation, except that the justification for the MSD formulation is intuitive, and does not have a direct decision theoretic justification.

The general MSD formulation that does not take advantage of the special structure of the binary variable classification problem is given as Problem IV:

Problem IV:

$$\text{Minimize } z_4 = \sum_{i=1}^{t} d_i$$

Subject to:

$$\mathbf{x}_i^{\mathrm{T}}\mathbf{w} - d_i \leq c, \qquad i \in G_1,$$
$$\mathbf{x}_i^{\mathrm{T}}\mathbf{w} + d_i > c, \qquad i \in G_2,$$
$$w_k \text{ and } c \text{ are unrestricted, } k = 1, \ldots, p,$$
$$d_i \geq 0, \qquad i = 1, \ldots, n,$$

Problem IV has n deviational variables and n proper constraints, one for each training sample observation. This formulation is the direct counterpart of Problem I, with the deviational variables d_i replacing the binary variables δ_i. As in Problem I, we can organize Problem IV by cell, combining all observations in cell $\mathcal{C}(s)$ to a single binary variable $\mathbf{b}_s$, yielding Problem V:

Problem V:
(Conventional Cell MSD)

$$\text{Minimize } z_5 = \sum_{s=1,\, \mathcal{C}(s) \neq \emptyset}^{t} n_{1s} d_{1s} + n_{2s} d_{2s}$$

Subject to:

$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w} - d_{1s} \leq c, \qquad \text{if } n_{1s} > 0,\ s = 1, \ldots, t,$$
$$\mathbf{b}_s^{\mathrm{T}}\mathbf{w} + d_{2s} > c, \qquad \text{if } n_{2s} > 0,\ s = 1, \ldots, t,$$
$$w_k \text{ and } c \text{ are unrestricted, } k = 1, \ldots, p,$$
$$d_{1s},\ d_{2s} \geq 0, \qquad s = 1, \ldots, t,$$

Problem V has up to $2t$ deviational variables and up to $2t$ constraints. The optimal solution to Problem V is identical to that of Problem IV. The Problem V formulation of the example problem introduced above is as follows:

$$\text{Minimize } z_5 = 15d_{11} + 6d_{21} + 8d_{12} + 17d_{22} + 6d_{13} + 12d_{23} + 21d_{14} + 5d_{24}$$

$$\begin{aligned}
0 \times w_1 + 0 \times w_2 - d_{11} &\leq c, && \text{for } \mathcal{C}(1),\\
0 \times w_1 + 0 \times w_2 + d_{21} &> c, && \text{for } \mathcal{C}(1),\\
0 \times w_1 + 1 \times w_2 - d_{12} &\leq c, && \text{for } \mathcal{C}(2),\\
0 \times w_1 + 1 \times w_2 + d_{22} &> c, && \text{for } \mathcal{C}(2),\\
1 \times w_1 + 0 \times w_2 - d_{13} &\leq c, && \text{for } \mathcal{C}(3),\\
1 \times w_1 + 0 \times w_2 + d_{23} &> c, && \text{for } \mathcal{C}(3),\\
1 \times w_1 + 1 \times w_2 - d_{14} &\leq c, && \text{for } \mathcal{C}(4),\\
1 \times w_1 + 1 \times w_2 + d_{24} &> c, && \text{for } \mathcal{C}(4),\\
w_1,\ w_2 &\text{ and } c \text{ are unrestricted},\\
d_{1s},\ d_{2s} &\geq 0,\ s = 1, \ldots, 4.
\end{aligned}$$

Development of Cell Reduced MSD Formulation

Similar to the BMIP formulation in Problem III, where $\delta_{1s}\delta_{2s} = 0$, it is easy to show that in Problem V, $d_{1s}d_{2s} = 0$, for each s. However, whereas in the BMIP formulation $\delta_{1s} + \delta_{2s} = 1$, in Problem V there is no general expression for $d_{1s} + d_{2s}$. Nevertheless, from a classification viewpoint we can limit ourselves to using only one value d_s, because either $\mathbf{b}_s^{\mathrm{T}}\mathbf{w} \leq c$ or $\mathbf{b}_s^{\mathrm{T}}\mathbf{w} > c$, so that we may use only one inequality for each cell, namely that for the class with the greatest number of observations for this cell. Thus, each cell has at most one constraint associated with it, either $\mathbf{b}_s^{\mathrm{T}}\mathbf{w} - d_s \leq c$, if $n_{1s} \geq n_{2s}$ and $n_{1s} > 0$, or $\mathbf{b}_s^{\mathrm{T}}\mathbf{w} + d_s > c$, if $n_{1s} < n_{2s}$. We simplify the Cell Reduced MSD formulation criterion z_6 in (4.6),

$$\text{minimize } z_6 = \sum_{s=1,\,\mathcal{C}(s)\neq\emptyset}^{t} \Big\{ |\, n_{1s} - n_{2s} \,|\, d_s + \min(n_{1s},\, n_{2s}) \Big\}, \tag{4.6}$$

by omitting the second term, as it is merely a constant.

Problem VI:
(Cell Reduced MSD)

$$\text{Minimize } z_7 = \sum_{s=1,\,\mathcal{C}(s)\neq\emptyset}^{t} |\, n_{1s} - n_{2s} \,|\, d_s$$

Subject to:

$$\begin{aligned}
\mathbf{b}_s^{\mathrm{T}}\mathbf{w} - d_s &\leq c, \quad \text{if } n_{1s} \geq n_{2s},\ n_{1s} > 0,\ s = 1, \ldots, t,\\
\mathbf{b}_s^{\mathrm{T}}\mathbf{w} + d_s &> c, \quad \text{if } n_{1s} < n_{2s},\ s = 1, \ldots, t,\\
w_k &\text{ and } c \text{ are unrestricted},\ k = 1, \ldots, p,\\
d_s &\geq 0, \quad i = 1, \ldots, t,
\end{aligned}$$

Problem VI has at most t deviational variables, and at most t proper constraints. The formulation in Problem VI does not have a decision theoretic justification. However, an intuitive motivation is that criterion z_7 weights the "balance of evidence" $|\, n_{1s} - n_{2s} \,|$ of the number of observations belonging to G_1 and G_2 in each cell $\mathcal{C}(s)$ by the undesirable distance d_s of $\mathbf{b}_s^{\mathrm{T}}\mathbf{w}$ from the surface separating the groups. The example problem formulation according to Problem VI is as follows:

$$
\begin{aligned}
&\text{Minimize } z_7 = 9d_1 + 9d_2 + 6d_3 + 16d_4 \\
&0 \times w_1 + 0 \times w_2 - d_1 \leq c, \qquad \text{for } \mathcal{C}(1), \\
&0 \times w_1 + 1 \times w_2 + d_2 > c, \qquad \text{for } \mathcal{C}(2), \\
&1 \times w_1 + 0 \times w_2 + d_3 > c, \qquad \text{for } \mathcal{C}(3), \\
&1 \times w_1 + 1 \times w_2 - d_4 \leq c, \qquad \text{for } \mathcal{C}(4), \\
&w_1,\ w_2 \text{ and } c \text{ are unrestricted}, \\
&d_s \geq 0,\ s = 1, \ldots, 4.
\end{aligned}
$$

5. EXAMPLES

We use three real data sets to illustrate the effectiveness of the MP approaches to binary variable classification. An advantage of using real data sets is that the classification results are not artificially biased in favor of certain methods. Of course, the use of real data also limits the scope of any conclusions. The MP methods were solved using LINDO (Schrage 1991). For the other methods we used our own programs. We used Hall's estimator of the smoothing parameters of the kernel procedure. All experiments were carried out on an IBM compatible PC 486/80 MHz.

5.1. Data Sets

Example 1

The data of the first example pertain to a study conducted to construct a prognostic index for predicting postoperative pulmonary embolism (PPE), using information on 395 patients who had surgery at the Military Medical Academy in Sofia. Of these patients, 141 developed PPE and 254 did not. Three binary variables were used to predict PPE. Each of these variables represents the presence or absence of a symptom of PPE, with each variable equal to 1 iff the symptom is present: x_1 indicates the presence of cancer as the main disease; x_2 the presence of at least one of the following moderate risk concomitant diseases – cardiac failure, local atherosclerosis, diabetes, hypertonia, varicisis, pulmonary emphysema; and x_3 the presence of at least one of the following high risk concomitant diseases – syndrome postphlebitic, cardiac decompensation, chronic lung disease, general atherosclerosis. The distribution of the patients over the multinomial cells is given in Table 2.

Table 2 About Here

Example 2

The data set of the second example consists of 242 patients at the National Center for Emergency Medicine in Bulgaria, 102 of whom were diagnosed with dissecting aneurysm (DA) and 140 were diagnosed with other, similar diseases (Other: 40 with pulmonary embolism, 50 with angina pectoris, and 50 with myocardial infarction). In our analysis, we seek to diagnose each patient as belonging to one of these groups (DA or Other), based on three symptoms: x_1, albuminuria; x_2, paroxysmal suffocation; and x_3, conscious disturbances. Each of these variables equals 1 if the symptom is present, and 0 if the symptom is absent. The actual patient distribution over the multinomial cells is given in Table 3.

Table 3 About Here

Example 3

The third data set contains information on 144 children who suffered from cranial trauma, collected at the Department of Pediatrics, Medical Faculty of Sofia. Of these children, 94 did not suffer from posttraumatic epilepsy (NO) and 50 did suffer from posttraumatic epilepsy (PE). Three binary variables were used to describe the symptoms: x_1, the presence or absence of seizures during the first month after the trauma; x_2, the presence or absence of previous psychoneurological disturbances; and x_3, the presence or absence of treatment immediately after the trauma. Again, x_i equals 1 if the corresponding symptom is present, and 0 if the symptom is absent. The multinomial table with the distribution of the training sample is given in Table 4.

Tables 4 and 5 About Here

5.2. Discussion of Results

For each data set, we can establish the minimum possible number of misclassifications by summing the values of $\min(n_{1s}, n_{2s})$ over each $\mathcal{C}(s)$. Therefore, the optimal solution of $\sum_s \min(n_{1s}, n_{2s})$ equals 81, 95 and 35 misclassifications for Data Sets 1, 2 and 3, respectively. The 19 different classification methods included in our study are listed in Table 5. We use two types of error measures to evaluate each classification method: the resubstitution error (RES), which measures the number of misclassifications in the training sample, and the leave-one-out (LOO) or cross-validation error (Lachenbruch and Mickey 1968). In the LOO method, the number of misclassifications is determined by removing one observation from the training sample, estimating the classification rule based on the remaining training sample observations, then classifying the observation that was held out, and repeating this process, holding each observation out successively.

From Table 5, we see that the Full Multinomial, kNN-Hall (order 1, 2, 3), LLM (order 2), Bahadur (order 2), BMIP and Cell Reduced MSD methods each obtained the optimal solution for all three data sets. Interestingly, the BMIP and Cell Reduced MSD methods not only yielded the optimal number of misclassifications, but also coincided fully in terms of the distribution of misclassifications over the different cells and the values of each w_k and c, for all three data sets. Since the linear BMIP and Cell Reduced MSD rules yielded the minimum number of misclassifications, there was no need to include nonlinear attribute terms.

The Kernel estimator, the Fourier procedure and the QLR achieve the optimal solution for two of the three data sets. The Fourier procedure performs poorly on Data Set 2. Both the first and second order kNN-Hills estimators gives poor classification results for all three data sets. The Cell Conventional MSD formulation of Problem V, the Bahadur, first order LLM, QDF and LDF models

yield solutions which are clearly inferior to the BMIP and Cell Reduced MSD for all three data sets, and the classification performance of the LR is inferior for two of the three data sets.

Of course it is well known that no one of the discriminant procedures is best in all cases, and the purpose of the limited comparative study in this paper is to illustrate the relative classification performance of various parametric and nonparametric statistical methods, and in particular the BMIP and Cell Reduced MSD.

6. CONCLUSIONS

We introduced a novel MIP formulation (BMIP) for solving the binary variable classification problem. We showed that the resulting classification rule not only has the usual geometric interpretation of other MP-based formulations, but also posesses a strong decision theoretical justification, as the resulting classification rule minimizes the total probability of misclassification. Additionally, the BMIP formulation requires substantially less binary variables than general MIP formulations, enabling the analysis of almost any size training sample. In comparing the classification accuracy of the BMIP with a number of the most widely used parametric and nonparametric statistical methods on three different real data sets, we found the BMIP to perform better than, for instance, the LDF, QDF, kNN-Hills, first order LLM, LR, QLR, first order Bahadur model, Fourier procedure, Kernel estimator, and the cell conventional MSD, and at least as well as the other methods considered. In each case, the BMIP achieved the minimum possible number of misclassifications, both using the resubstitution and the leave-one-out error measures.

The current research can be extended in several different ways. First, additional comparative studies are needed to further establish the classificatory performance of the BMIP formulation. Second, it is of interest to develop decision theoretic MP-based formulations based on equations (2.1) and (2.3). Third, it appears useful to explore decision theoretic MP formulations based on variants of the MSD criterion. Fourth, future research should analyze the extension of the binary variable case to that of mixed variables and general categorical variables.

REFERENCES

Agresti, A., *Categorical Data Analysis*, Wiley, New York, NY, 1990.

Aitchison, J. and Aitken, C. G. G., "Multivariate Binary Discrimination by the Kernel Method," *Biometrika*, **63**, 1976, 413–420.

Anderson, J. A., "Separate Sample Logistic Discrimination," *Biometrika*, **59**, 1972, 19–35.

Anderson, J. A., "Quadratic Logistic Discrimination," *Biometrika*, **62**, 1975, 149–154.

Anderson, T. W., *An Introduction to Multivariate Statistical Analysis, Second Edition*, Wiley, New York, NY, 1984.

Anderson, J. A., Whaley, K., Williamson, J. and Buchanan, W. W., "A Statistical Aid to the Diagnosis of Keratoconjunctivitis Sicca," *Quarterly Journal of Medicine*, **41**, 1972, 175–189.

Asparoukhov, O. K., *Microprocessor System for Investigation of Thromboembolic Complications*, Unpublished Ph.D. Dissertation, Technical University of Sofia, Bulgaria (in Bulgarian), 1985.

Asparoukhov, O. K. and Andreev, T. B., "Comparison of One-Stage Classifiers for Assessment of the Ability of Children to Construct Grammatical Structures Consciously," in *Multivariate Analysis in the Behavioral Sciences. Philosophical to Technical*, I. Panchev (Ed.), Academic Publishing House 'Prof. Marin Drinov,' Sofia, Bulgaria, 1995, 1–13.

Bahadur, R. R., "A Representation of the Joint Distribution of Response to n Dichotomous Items," in *Studies in Item Analysis and Prediction*, H. Solomon (Ed.), Stanford University Press, Palo Alto, CA, 1961, 158–168.

Banks, W. J. and Abad, P. L., "An Efficient Optimal Solution Algorithm for the Classification Problem," *Decision Sciences*, **22**, 1991, 1008–1023.

Banks, W. J. and Abad, P. L., "On the Performance of Linear Programming Heuristics Applied on a Quadratic Transformation in the Classification Problem," *European Journal of Operational Research*, **74**, 1994, 23–28.

Benediktson, J. A. and Swain, P. H., "Consensus Theoretic Classification Methods," *IEEE Transactions on Systems, Man and Cybernetics*, **22**, 1992, 688–704.

Cox, D. R., "Some Procedures Connected with the Logistic Qualitative Response Curve," in *Research Papers in Statistics: Festschrift for J. Neyman*, F. N. David (Ed.), Wiley, New York, 1966, 55–71.

Das Gupta, S., "Theories and Methods in Classification: A Review," in *Discriminant Analysis and Applications*, T. Cacoullos (Ed.), Academic Press, New York, NY, 1973, 77–137.

Day, N. E. and Kerridge, D. F., "A General Maximum Likelihood Discriminant," *Biometrics*, **23**, 1967, 313–323.

Dillon, W. R. and Goldstein, M., "On the Performance of Some Multinomial Classification Rules," *Journal of the American Statistical Association*, **73**, 1978, 305–313.

Duarte Silva, A. P., *Minimizing Misclassification Costs in Two-Group Classification Analysis*, Unpublished Ph.D. Dissertation, The University of Georgia, 1995.

Duarte Silva, A. P. and Stam, A., "Second Order Mathematical Programming Formulations for Discriminant Analysis," *European Journal of Operational Research*, **72**, 1994, 4–22.

Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, **7**, 1936, 179–188.

Freed, N. and Glover, F., "A Linear Programming Approach to the Discriminant Problem," *Decision Sciences*, **12**, 68–74, 1981a.

Freed, N. and Glover, F., "Simple But Powerful Goal Programming Models for Discriminant Problems," *European Journal of Operational Research*, **7**, 1981b, 44–60.

Gehrlein, W. V., "General Mathematical Programming Formulation for the Statistical Classification Problem," *Operations Research Letters*, **5**, 1986, 299–304.

Gilbert, E. S., "On Discrimination Using Qualitative Variables," *Journal of the American Statistical Association*, **63**, 1968, 1399–1412.

Glorfeld, L. W. and Kattan, M. W., "A Comparison of the Performance of Three Classification Procedures When Applied to Contaminated Data," in *Proceedings of the 21th Annual Meeting of the Decision Sciences Institute*, 1989, 1153–1155.

Glover, F., Keene, S. and Duea, B., "A New Class of Models for the Discriminant Problem," *Decision Sciences*, **19**, 1988, 269–280.

Gochet, W., Stam, A., Srinivasan, V. and Chen, S., "Multi-Group Discriminant Analysis Using Linear Programming," *Operations Research*, Forthcoming, 1996.

Goldstein, M. and Dillon, W. R., *Discrete Discriminant Analysis*, Wiley, New York, NY, 1978.

Hall, P., "On Nonparametric Multivariate Binary Discrimination," *Biometrika*, **68**, 1981a, 287–294.

Hall, P., "Optimal Near Neighbour Estimator for Use in Discriminant Analysis," *Biometrika*, **68**, 1981b, 572–575.

Hand, D. J., *Discrimination and Classification*, Wiley, New York, NY, 1981.

Hand, D. J., "Statistical Pattern Recognition on Binary Variables," in *Pattern Recognition Theory and Applications*, J. Kittler, K. S. Fu and L. F. Pau (Eds.), Reidel, Boston, MA, 1982, 19–33.

Hand, D. J., "A Comparison of Two Methods of Discriminant Analysis Applied to Binary Data," *Biometrics*, **39**, 1983, 683–694.

Hand, D. J., "Discriminant Analysis for Categorical Data," in *Lecture Notes and Program of the 4th European Courses in Advanced Statistics Program: Analysis of Categorical Data Theory and Application*, Leiden, The Netherlands, 1993, 135–174.

Hartigan, J. A., *Clustering Algorithms*, Wiley, New York, NY, 1975.

Hills, M., "Discrimination and Allocation with Discrete Data," *Applied Statistics*, **16**, 1967, 237–250.

Ibaraki, T. and Muroga, S., "Adaptive Linear Classifier by Linear Programming," *IEEE Transactions on System Science and Cybernetics*, **SSC-6**, 1970, 53–62.

Joachimsthaler, E. A. and Stam, A., "Four Approaches to the Classification Problem in Discriminant Analysis: An Experimental Study," *Decision Sciences*, **19**, 1988, 322–333.

Joachimsthaler, E. A. and Stam, A., "Mathematical Programming Approaches for the Classification Problem in Two-Group Discriminant Analysis," *Multivariate Behavioral Research*, **25**, 1990, 427–454.

Koehler, G. J. and Erenguc, S. S., "Minimizing Misclassifications in Linear Discriminant Analysis," *Decision Sciences*, **21**, 1990, 63–85.

Krzanowski, W. J., "The Performance of Fisher's Linear Discriminant Function Under Non-Optimal Conditions," *Technometrics*, **19**, 1977, 191–200.

Krzanowski, W. J., "Distance Between Populations Using Mixed Continuous and Categorical Variables," *Biometrika*, **70**, 1983, 235–243.

Krzanowski, W. J., *Principles of Multivariate Analysis*, Clarendon Press, Oxford, England, 1988.

Krzanowski, W. J., "The Location Model for Mixture of Categorical and Continuous Variables," *Journal of Classification*, **10**, 1993, 25–49.

Lachenbruch, P. A. and Mickey, M. R., "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, **10**, 1968, 1–11.

Lam, K. F., Choo, E. U. and Wedley, W. C., "Linear Goal Programming in Estimation of Classification Probability," *European Journal of Operational Research*, **67**, 1993, 101–110.

Liitschwager, J. M. and Wang, C. "Integer Programming Solution of a Classification Problem," *Management Science*, **24**, 1978, 1515–1525.

Lippmann, R. P., "Pattern Classification Using Neural Networks," *IEEE Communication Magazine*, 1989, 47–64.

Martin, D. C. and Bradly, R. A., "Probability Models, Estimation, and Classification for Multivariate Dichotomous Populations," *Biometrics*, **28**, 1972, 203–221.

Másson, E. and Wang, Y.-J., "Introduction to Computation and Learning in Artificial Neural Networks," *European Journal of Operational Research*, **47**, 1990, 1–28.

McLachlan, G. J., *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, NY, 1992.

Mangasarian, O. L., "Linear and Nonlinear Separation of Patterns by Linear Programming," *Operations Research*, **13**, 1965, 444–452.

Markowski, C. A. and Markowski, E. P., "An Experimental Comparison of the Discriminant Problem with Both Qualitative and Quantitative Variables," *European Journal of Operational Research*, **28**, 1987, 74–78.

Minnick, R. C., "Linear-Input Logic," *IEEE Transactions on Electronics and Computers*, **EC–10**, 1961, 6–16.

Moore, D. H., "Evaluation of Five Discriminant Procedures for Binary Variables," *Journal of the American Statistical Association*, **68**, 1973, 399–404.

Ng, K.-C. and Abramson, B., "Consensus Diagnosis: A Simulation Study," *IEEE Transactions on Systems, Man and Cybernetics*, **22**, 1992, 916–928.

Ott, J. and Kronmal, R. A., "Some Classification Procedures for Multivariate Binary Data Using Orthogonal Functions," *Journal of the American Statistical Association*, **71**, 1976, 391–399.

Ragsdale, C. T. and Stam, A., "Introducing Discriminant Analysis to the Business Statistics Curriculum," *Decision Sciences*, **23**, 1992, 724–745.

Rubin, P. A., "Evaluating the Maximize Minimum Distance Formulation to the Linear Discriminant Problem," *European Journal of Operational Research*, **41**, 1989, 272–282.

Rubin, P. A., "A Comment Regarding Polynomial Discriminant Analysis," *European Journal of Operational Research*, **72**, 1994, 29–31.

Rypley, B., "Neural Networks and Related Methods for Classification," *Journal of the Royal Statistical Society*, **B**, **56**, 1994, 409–456.

Salchenberger, L. M., Cinar, E. M. and Lash, N. A., "Neural Networks: A New Tool for Predicting Thrift Failures," *Decision Sciences*, **23**, 1992, 899–916.

Schrage, L., *LINDO: User's Manual, Release 5.0*, The Scientific Press, South San Francisco, CA, 1991.

Smith, C. A. B., "Some Examples of Discrimination," *Annals of Eugenics*, **13**, 1947, 272–282.

Smith, F. W., "Pattern Classifier Design by Linear Programming," *IEEE Transactions on Computing*, C–17, 1968, 367–372.

Soltysik, R. C. and Yarnold, P. R., *ODA 1.0: Optimal Discriminant Analysis for DOS*, Optimal Data Analysis, Chicago, IL, 1993.

Soltysik, R. C. and Yarnold, P. R., "The Warmack-Gonzalez Algorithm for Linear Two-Category Multivariate Optimal Discriminant Analysis," *Computers & Operations Research*, **21**, 1994, 735–745.

Stam, A. and Joachimsthaler, E. A., "Solving the Classification Problem in Discriminant Analysis via Linear and Nonlinear Programming Methods," *Decision Sciences*, **20**, 1989, 285–293.

Stam, A. and Joachimsthaler, E. A., "A Comparison of a Robust Mixed-Integer Approach to Existing Methods for Establishing Classification Rules for the Discriminant Problem," *European Journal of Operational Research*, **46**, 1990, 113–122.

Stam, A. and Jones, D. G., "Classification Performance of Mathematical Programming Techniques in Discriminant Analysis: Results for Small and Medium Sample Sizes," *Managerial and Decision Economics*, **11**, 1990, 243–253.

Sturt, E., "Algorithm AS 165. An Algorithm to Construct a Discriminant Function in Fortran for Categorical Data," *Applied Statistics*, **30**, 1980, 313–325.

Tam, K. Y. and Kiang, M. Y., "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," *Management Science*, **38**, 1992, 926–947.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. and Gelpke, G. J., "Comparison of Discriminant Techniques Applied to a Complex Data Set of Head Injured Patients" (with discussion), *Journal of the Royal Statistical Society, Series A*, **144**, 1981, 145–175.

Trampisch, H. J., "Classical Discriminant Analysis and Lancaster Models for Qualitative Data," in *Compstat 1978, Proceedings in Computational Statistics*, Physica-Verlag, Vienna, 1978, 205–211.

Wanarat, P. and Pavur, R., "Examining the Effect of Second-Order Terms in Mathematical Programming Approaches to the Classification Problem," *European Journal of Operational Research*, Forthcoming, 1996.

Xu, L., Krzyzak, A. and Suen, C. Y., "Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition," *IEEE Transactions on Systems, Man and Cybernetics*, **22**, 1992, 418–435.

Yarnold, P. R., Soltysik, R. C. and Martin, G. J., "Heart Rate Variability and Susceptibility for Sudden Cardiac Death: An Example of Multivariable Optimal Discriminant Analysis," *Statistics in Medicine*, **13**, 1994, 1015–1021.

Young, T. Y., Liu, P. S. and Rondon, R. J., "Statistical Pattern Classification with Binary Variables," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 1981, 155–163.

Table 1: Example Problem

Cell $\mathcal{C}(s)$	Attribute Values		Number of Observations		$\lvert n_{1s}-n_{2s} \rvert$	$\text{Min}(n_{1s}, n_{2s})$
s	x_1	x_2	G_1	G_2		
1	0	0	15	6	9	6
2	0	1	8	17	9	8
3	1	0	6	12	6	6
4	1	0	21	5	16	5

Table 2: Training Sample Distribution, Data Set 1[1]

Cell $\mathcal{C}(s)$	Attribute Values			Number of Patients	
s	x_1	x_2	x_3	PPE[2]	NO[3]
1	0	0	0	8	74
2	0	0	1	30	123
3	0	1	0	15	29
4	0	1	1	23	18
5	1	0	0	17	3
6	1	0	1	38	7
7	1	1	0	3	0
8	1	1	1	7	0

1: $n_1 = 141$, $n_2 = 254$, $n = 395$.
2: Patients with postoperative pulmonary embolism.
3: Patients with no postoperative pulmonary embolism.

Table 3: Training Sample Distribution, Data Set 2[1]

Cell $\mathcal{C}(s)$	Attribute Values			Number of Patients	
s	x_1	x_2	x_3	DA[2]	Others[3]
1	0	0	0	34	39
2	0	0	1	8	14
3	0	1	0	17	15
4	0	1	1	5	2
5	1	0	0	19	28
6	1	0	1	3	26
7	1	1	0	12	10
8	1	1	1	4	6

1: $n_1 = 102$, $n_2 = 140$, $n = 242$.
2: Patients with dissecting aneurism.
3: Patients with other diseases: pulmonary embolism, angina pectoris, myocardial infarction.

Table 4: Training Sample Distribution, Data Set 3[1]

Cell $\mathcal{C}(s)$	Attribute Values			Number of Patients	
s	x_1	x_2	x_3	NO[2]	PE[3]
1	0	0	0	16	10
2	0	0	1	0	5
3	0	1	0	12	7
4	0	1	1	1	6
5	1	0	0	39	8
6	1	0	1	5	3
7	1	1	0	21	6
8	1	1	1	0	5

1: $n_1 = 94$, $n_2 = 50$, $n = 144$.
2: Patients with no posttraumatic epilepsy.
3: Patients with posttraumatic epilepsy.

Table 5: Number and Percentage of Misclassified Observations, 19 Classification Methods

Procedure		Number Misclassified						Percentage Misclassified					
		Data Set 1		Data Set 2		Data Set 3		Data Set 1		Data Set 2		Data Set 3	
		RES[1]	LOO[2]	RES	LOO	RES	LOO	RES	LOO	RES	LOO	RES	LOO
Full Multinomial		81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
kNN-Hills	L=1	131	131	100	100	40	40	33.2	33.2	41.3	41.3	27.8	27.8
	L=2	141	141	102	102	50	50	35.7	35.7	42.2	42.2	34.7	34.7
kNN-Hall	L=1	81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
	L=2	81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
	L=3	81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
LLM	Order 1	86	86	98	98	37	37	21.8	21.8	40.5	40.5	25.7	25.7
	Order 2	81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
Bahadur Model	Order 1	86	86	98	98	37	37	21.8	21.8	40.5	40.5	25.7	25.7
	Order 2	81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
Kernel Estimator		81	81	95	95	37	37	20.5	20.5	39.3	39.3	25.7	25.7
Fourier Procedure		81	81	95	124	35	35	20.5	20.5	39.3	51.2	24.3	24.3
LDF		86	104	98	98	37	37	21.8	26.3	40.5	40.5	25.7	25.7
QDF		81	104	95	95	37	37	20.5	26.3	39.3	39.3	25.7	25.7
LR		81	81	98	98	37	37	20.5	20.5	40.5	40.5	25.7	25.7
QLR		81	81	95	95	35	40	20.5	20.5	39.3	39.3	24.3	27.8
MIP		81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
MSD (Cell Reduced)		81	81	95	95	35	35	20.5	20.5	39.3	39.3	24.3	24.3
MSD (Cell Conventional)		86	86	97	97	37	37	21.8	21.8	40.1	40.1	25.7	25.7
Optimal Number of Misclassifications		81		95		35		20.5		39.3		24.3	

1: RES = Using the resubstitution method.
2: LOO = Using the leave-one-out method.