THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

Date July 28, 2009

WP # 0102MSS-61-2009

A Mixed Integer Programming Model for Multiple-Class Discriminant Analysis

Minghe Sun

Department of Management Science and Statistics
The University of Texas at San Antonio

UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631
210 458-4317 | BUSINESS.UTSA.EDU

A Mixed Integer Programming Model for Multiple-Class Discriminant Analysis

Minghe Sun

Department of Management Science and Statistics
College of Business
The University of Texas at San Antonio
San Antonio, TX 78249-0632
(210) 458-5777 (phone), (210) 458-6350 (fax)
msun@utsa.edu
http://business.utsa.edu/faculty/msun

A Mixed Integer Programming Model for Multiple-Class Discriminant Analysis

Abstract

A mixed integer programming model is proposed for multiple-class discriminant and classification analysis. When multiple discriminant functions, one for each class, are constructed with the mixed integer programming model, the number of misclassified observations in the sample is minimized. Although having its own right, this model may be considered as a generalization of mixed integer programming formulations for two-class classification analysis. Properties of the model are studied. The model is immune from any difficulties of many mathematical programming formulations for two-class classification analysis, such as nonexistence of optimal solutions, improper solutions and instability under linear data transformation. In addition, meaningful discriminant functions can be generated under conditions other techniques fail. Results on data sets from the literature and on data sets randomly generated show that this model is very effective in generating powerful discriminant functions.

Keywords: **Discriminant Analysis; Classification; Mixed Integer Programming; Optimization; Nonparametric Procedures**

JEL Codes: **C14, C61**

A Mixed Integer Programming Model for Multiple-Class Discriminant Analysis

1. Introduction

Although extensive studies have been undertaken in mathematical programming (MP) approaches for discriminant and classification analysis, the focus has been on two-class classification techniques. Generalizations to multiple-class techniques have been attempted, but researchers are not completely satisfied with these earlier generalizations. A mixed integer programming (MIP) model is proposed in this study as a nonparametric procedure for multiple-class discriminant and classification analysis. Although it is an extension of the linear programming (LP) model of Sun [2002b], this MIP model may be considered as a generalization of the MIP models for two-class classification analysis.

Sun [2002b] proposed a simple but powerful LP model for this purpose. The LP model minimizes the sum of deviations of misclassified observations in the sample, or the L_1 -norm. The MIP model proposed in this study minimizes the number of misclassified observations in the sample, or the L_0 -norm. The LP model of Sun [2002b] has very good properties, is immune to the pathologies of many other earlier MP models and should work well under all situations. However, a MIP model is appealing because it directly minimizes the number of misclassified observations in the sample and some authors have reported that some MIP models for two-class classification outperformed other models under certain conditions. In addition, a MIP model should be always the choice if the purpose of the application is discrimination rather than classification.

Some properties of the MIP model are studied. Like the LP model [Sun, 2002b], the MIP model is immune to the difficulties caused by pathologies of earlier MP models for two-class classification analysis. For decades, research in MP approaches for discriminant and classification analysis has been focused on the two-class problems. Researchers have spent many years looking for simple but powerful generalizations from the two-class techniques to multiple-class techniques. The MIP model proposed in this study provides such simple but powerful generalization and may make MP approaches attractive and better alternatives to other discriminant techniques.

Discriminant and classification analysis has been fundamental scientific research and practical applications over many decades. Discriminant analysis involves the study of the differences between two or more classes of objects that are described by measurements, or prediction variables, of different characteristics or attributes. Classification involves the study of assigning new observations into one of the two or more classes based on the measurements on the different characteristics. Applications of discriminant and classification analysis are diverse. To mention a few, applications in business include financial management [Alman, 1968; Srinivasan and Kim, 1987; Zopounidis, 1998; Zopounidis and Dimitras, 1998], human resource management [Rulon, Tiedeman, Tatsuoka and Langmuir, 1967; Walker, 1974], marketing [Dutka, 1995], student recruiting [Choo and Wedley, 1985]; applications in biology and medicine include patient classification [Happer, 2005], disease diagnosis [Dudoit, Fridlyand and Speed, 2002; Sun and Xiong, 2002] and species classification [Fisher, 1936]; and applications in environment and geography include remote sensing image pattern classification [Shankar, Meher, Ghosh and Bruzzone, 2007; Yin

and Guo, 2007] and pollution control [Rossi, Slowinski and Susmanga, 1999], among others. In fact, this study was motivated by the need to identify a few from many thousands of genes that can be used to classify tissue samples into normal and tumor tissues and to identify genes responsible for certain diseases [Sun and Xiong, 2002a, 2002b]. Discriminant and classification analysis techniques will play a more significant role in data analysis as information technology advances and as huge amount of data need to be analyzed with data mining tools. The availability of large sets of data collected through information technology such as the Internet, imaging and spectrome, made the traditional discriminant and classification techniques inadequate.

Based on known values of the attributes or variables and known class memberships of the observations in a sample, usually called a training sample, mathematical discriminant functions are constructed. The attribute values of an observation can be evaluated by these discriminant functions to obtain discriminant scores and the observation is assigned to a class based on these discriminant scores. For many decades, statistical techniques, such as Fisher's linear discriminant function (LDF) [Fisher, 1936], Smith's quadratic discriminant function (QDF) [Smith, 1947] and logistics regression [Hand, 1981], have been standard tools for this discriminant and classification analysis. Statistical methods perform well when the data analyzed satisfy the underlying assumptions, such as multivariate normality and equal covariance matrices of the prediction variables, although minor deviations from these assumptions do not severely affect the performance of these statistical methods. More recently, other techniques, such as MP [Freed and Glover, 1981a, 1981b; Hand, 1981] including support vector machines (SVM) [Vapnik, 1995, 1998], neural networks [Stern, 1996], and classification trees [Breiman, Friedman, Olshen and Stone, 1984] have become alternative tools for discriminant and classification analysis. A spectrum of techniques is needed because no single technique always outperforms others under all situations.

2. MP Approaches to Discriminant Analysis

The publication of the original LP models for two-class classification [Freed and Glover, 1981a; Hand, 1981] inspired a series of studies. A considerable number of publications on this topic have appeared in the literature. Some of these articles reported limitations and pathologies of some of the earlier MP models, some provided diagnoses, and others offered remedies to improve the earlier models resulting alternative or improved MP models [Cavalier, Ignizio and Soyster, 1989; Freed and Glover, 1986b; Glover, 1990; Glover, Keene, and Duea 1988; Koehler, 1989a, 1989b, 1990, 1991; Markowski and Markowski, 1985]. The different MP models introduced in the literature include LP, MIP, goal programming, nonlinear programming and quadratic programming (*i.e.*, SVM) models [Erenguc and Koehler, 1990; Stam, 1997; Stam and Joachimsthaler, 1990]. Through these studies, the MP techniques, especially for the two-class discriminant and classification analysis, are maturing quickly.

Three difficulties due to pathologies of some earlier MP formulations have caused concerns [Freed and Glover, 1986b; Koehler, 1989a, 1989b, 1990, 1991; Markowski and Markowski, 1985]. These difficulties are unbounded objective function, degenerate or improper solutions, and solution instability under linear data transformation. The objective function of a MP formulation is unbounded if its value can be made arbitrarily large for a maximization problem or arbitrarily small for a minimization problem resulting in no meaningful solutions. A solution is degenerate or trivial if all the estimated coefficients in the classification function or in the discriminant

functions are 0. A solution is improper if all the resulting discriminant functions for different classes are the same and therefore none of the observations can be definitely assigned to any class. A solution is proper if it is not improper. Solution instability occurs when a MP formulation produces different sets of discriminant functions with different classification results when the data are linearly transformed. As will be shown later, the MIP model proposed in this study is immune from these difficulties. The diagnoses include the conditions under which these difficulties may occur. The remedies and improvements include normalizations of the coefficients in classification functions to avoid unbounded solution, trivial solutions and improper solutions.

The concept of a trivial solution in the context of multiple-class classification is different from that in two-class classification. In two-class classification analysis, a classification function representing a hyperplane separating the two classes is constructed. In multiple-class classification analysis, p discriminant functions are constructed, one for each class, and the hyperplane separating two classes is where the two discriminant functions have the same discriminant scores. Therefore, the equation representing the hyperplane is where the difference of the two discriminant functions is 0. As a result, when all discriminant functions have the same coefficients, *i.e.*, an improper solution, the equations representing the boundaries of any two classes all have coefficients of 0's, *i.e.*, a degenerate solution. In multiple-class discriminant analysis, improper solutions may occur and the trivial solution becomes a special case of improper solutions.

Unlike statistical methods, the MP approaches, as nonparametric methods, do not make strict assumptions about the data analyzed. Many studies comparing the performances between the more traditional statistical methods, such as Fisher's LDF and Smith's QDF, and MP approaches have been reported [Freed and Glover, 1986a; Nath, Jackson and Jones, 1992; Joachimsthaler and Stam, 1988]. Many computational experiments have been undertaken [Bajaier and Hill, 1982; Freed and Glover, 1986a; Joachimsthaler and Stam, 1988; Markowski and Markowski, 1987; Rubin, 1989b, 1990b; Stam and Joachimsthaler, 1990]. Some of these studies also compared the performances of different mathematical formulations. Some of these studies used real data and others used simulated data. In general, the conclusion is that no single technique performs the best under all conditions. MP approaches outperform statistical approaches when the assumptions underlying the statistical approaches are seriously violated [Ragsdale and Stam, 1991; Stam and Joachimsthaler, 1990]. Being able to perform well on a variety of types of data is an advantage of MP approaches. Another advantage of MP approaches over the traditional statistical techniques is that the fitted model is less influenced by outlier observations. Nath and Jones [1988], Glen [1999, 2001] and Sun and Xiong [2002a, 2002b] have addressed the problem of variable selection in MP approaches for discriminant analysis.

Although most of the research in MP approaches is around two-class classification, attempts have been made to extend the approaches to multiple-class discriminant and classification analysis [Choo and Wedley, 1985; Freed and Glover 1981b; Gehrlein, 1986; Gochet, Stam, Srinivasan and Chen, 1977; Pavur, 1997; Pavur and Loucopoulos, 1995; Sun, 2002]. Each of these models has its merits although each has drawbacks [Gochet, Stam, Srinivasan and Chen, 1977; Pavur and Loucopoulos, 1995; Stam, 1997]. In general, the research community is not fully satisfied with these models [Stam, 1997]. Without a simple but powerful generalization, the MP approaches are handicapped and will never be able to compete with other techniques. Researchers and practitioners will be

more willing to accept the MP approaches as nonpapamatric procedures when simple but powerful multiple-class MP models are available. Given the difficulty of multiple-class classification problems, some researchers focused on the three-class classification problem [Loucopoulos, 2001; Loucopoulos and Pavur, 1997a, 1997b; Pavur and Loucopoulos, 2001].

Conceiving that an extension from techniques for two-class classification to those for multiple-class classification was straightforward, Freed and Glover [1981b] proposed a decomposition of a p-class discriminant problem into $p(p-1)/2$ two-class discriminant problems. Each problem represents a pair of classes and each is solved separately to determine a classification function representing the hyperplane separating the two classes. This approach is later on called the one-against-one approach in the literature. The drawback of this approach is that the classification functions may be sub-optimal because these functions are not estimated in an aggregate form. As a result, the classification of observations in some segments of the variable space is not clear [Loucopoulos and Pavur, 1997a; Pavur and Loucopoulos, 1995; Stam, 1997]. Furthermore, this approach is tedious because too many subproblems are formulated and solved and too many classification functions are estimated.

Another extension is to decompose a p-class discriminant problem into a p two-class problem. In the k th two-class problem, the observations in class k are treated as one class and all the rest are treated as the other. This approach is later on called the one-against-all approach [Vapnik, 1995, 1007]. This approach has the same drawbacks as those of the one-against-one approach except that p, instead of $p(p-1)/2$, classification functions are estimated. The advantage of this approach is that the p two-class problems are computationally easier to solve than a multiple-class model.

Gehrlein [1986] proposed two MIP formulations. These MIP models set up the foundations of most of the later studies in this area. One MIP model uses a single discriminant function with class specific cutoff discriminant scores and is referred to as the single function model. A new observation is assigned to a specific class if its evaluated value, *i.e.*, its discriminant score, falls into the interval for this class. This model implicitly implies that the hyperplanes separating the classes are all in parallel and the variable space is cut into layers by these hyperplanes. However, this is rarely the case for practical applications as shown by scatter plots of some of the prediction variables. Therefore, this model has pathological problems as conceived by many researchers [Stam, 1997]. Pavur and Loucopoulos [1995] modified the original single function model of Gehrlein [1986] from the minimization of the number of misclassified cases in the sample to the minimization of the sum of deviations of misclassified cases. This modified model, as a LP model, saved computation time for one randomly generated example problem. Östermark and Höglund [1998] attributed the single function multiple-class model of Gehrlein [1986] to Freed and Glover [1981b] and pointed out that the ordering of the classes are important for accurate classification because the class specific cutoff scores require that the class 1 scores be lower than class 2 scores that in turn be lower than class 3 scores, and so on. Therefore, they suggested the investigation of alternative sequencing of the classes. Choo and Wedley [1985] used multiple criterion decision making techniques to determine the coefficients in a single classification function for multiple-class discriminant problems.

Gehrlein [1986] also proposed a multiple discriminant function MIP model, one for each class, in a way analogous to statistical classification techniques. A new observation is evaluated by each discriminant function and is assigned to the class with the highest discriminant score. All the discriminant functions are estimated simultaneously and the model is pathologically correct. As a drawback, this MIP model requires a considerable number of binary variables and is computationally infeasible for problems with medium to large samples [Stam, 1997]. Because of its computational limitation, not many further studies on this model have been reported. Wilson [1996] introduced a multiple function MIP model as an alternative to the one in Gehrlein [1986]. In this MIP model, each observation is represented by $2p$ constraints and is associated with p binary variables. Therefore, this model is even more complicated and difficult to solve.

The model proposed by Bennet and Mangasarian [1994] is almost identical in structure to the multiple function model proposed by Gehrlein [1986]. However, instead of minimizing the L_0 -norm, Bennet and Mangasarian [1994] minimized a weighted L_1 -norm making the MP model computationally much easier to solve. The weight assigned to each term, representing the deviation if an observation is misclassified, in the objective function is the reciprocal of the sample size of the class that the observation belongs.

Gochet, Stam, Srinivasan and Chen [1997] proposed a LP model for multiple-class discriminant analysis. The objective function of this model is similar to but slightly different from that in Bennet and Mangasarian [1994]. The difference is that the terms in the objective function are not weighted. In addition to the same set of constraints as in Bennet and Mangasarian [1994], a normalization constraint is used restricting the difference between the sum of goodness of fit and the sum of badness of fit to be positive. Intuitively, a goodness of fit of an observation is the distance of the observation from the hyperplane separating the class to which the observation belongs and another class when the observation falls on the right side. A badness of fit is the same distance but when the observation falls on the wrong side. As in the statistical techniques, multiple discriminant functions are used, one for each class, and a new observation is assigned to the class with the highest discriminant score. All the discriminant functions are estimated simultaneously and the model is theoretically correct. Because no integer variables are involved in the formulation, the model is computationally very efficient to solve. However, this model is hard to implement without a special purpose software [Stam, 1997]. Östermark and Höglund [1998] extended the model of Gochet, Stam, Srinivasan and Chen [1997] to include quadratic terms of the variables in the discriminant functions.

Although it is possible to design computational experiments with randomly generated test problems with each class falling into layers in the variable space and to obtain good classification results, single function models are pathologically flawed and has very limited use in practice. Such single function models may perform well on certain data sets or under certain condition, but are not general to handle actual real life problems. Models using multiple discriminant functions, one for each class, are more general and flexible, are analogous to statistical techniques, and therefore are more preferred than single function models [Stam, 1997].

MIP approaches are greedier than many other methods, such as the LP approach or the Fisher's LDF. Therefore, it is possible for the MIP approach to achieve higher in-sample classification rate but lower validation classification rate than other methods. All MIP formulations of discriminant analysis have the major drawback of requiring excessive amount of computation time to solve. Their major attractive feature is that the number of

misclassified cases in the sample can be directly minimized. If the purpose of study is discrimination rather than classification of new observations, MIP models perform better than other models of similar structure.

3. Model Development

Assume there are a total of $p \geq 2$ known classes and $K = \{1, ..., p\}$ is used to denote the index set of all classes. A sample of m observations or cases is available and the class membership of each observation is unique and known. Among the m observations, m_k are form class k, for each $k \in K$, such that $m = \sum_{k \in K} m_k$. Let I denote the index set of all observations and I_k denote the index set of those from class k, for each $k \in K$, in the sample. The observations represent the objects to be classified. Assume n characteristics or prediction variables are used to describe the observations. The index set of all prediction variables is denoted by $J = \{1, ..., n\}$. The value of prediction variable $j \in J$ on observation i is denoted by x_{ij}. With $x_{i0} \equiv 1$ for all $i \in I$, $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{in})^{\mathrm{T}}$ represents the realized values of all prediction variables of observation i. Sometimes, $\mathbf{x} = (x_0, x_1, ..., x_n)^{\mathrm{T}}$ is used to denote the prediction variables of a generic observation. Some of the prediction variables are real and others may be nominal or categorical [Sun, 2002a]. When nonlinear discriminant functions are constructed, some of the x_j's measure the characteristics of the observations and, therefore, are independent variables, and others may be functions, such as squares or cross products, of other independent variables. When nonlinear discriminant functions are constructed, the MIP model proposed in this study is still linear because the functions are linear in the parameters although nonlinear in the variables.

It is assumed that there is a vector of unknown parameters $\boldsymbol{\beta}_k \in \Re^{n+1}$ with $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \cdots, \beta_{kn})$ for each $k \in K$ such that $\boldsymbol{\beta}_k \mathbf{x}$ is the discriminant function for class k. The elements of $\boldsymbol{\beta}_k$ need to be estimated using the data in the sample in such a way that observations in the sample are optimally classified according to a certain criterion. The sample estimate of $\boldsymbol{\beta}_k$ is denoted by $\mathbf{b}_k = (b_{k0}, b_{k1}, \cdots, b_{kn})$. The estimated discriminant functions $\mathbf{b}_k \mathbf{x}$ for $k \in K$ are then used to evaluate and classify observations. For any observation represented by $\mathbf{x} \in \Re^{n+1}$, discriminant scores are computed, *i.e.*,

$$g_k(\mathbf{x}) = \mathbf{b}_k \mathbf{x}_i \text{, for } k \in K. \tag{1}$$

The observation is assigned to a class k with the highest discriminant score, *i.e.*,

$$g_k(\mathbf{x}) = max\{g_{k'}(\mathbf{x}) \mid k' \in K\}. \tag{2}$$

Let $M > 0$ be a constant that is sufficiently large. For each observation $i \in I_k$, there always exist c_i and d_i such that the following p inequalities hold

$$\mathbf{b}_k \mathbf{x}_i + M d_i > c_i \tag{3}$$

$$\mathbf{b}_{k'} \mathbf{x}_i \leq c_i \qquad \text{for } k' \neq k, \tag{4}$$

where, c_i is a cutoff point and d_i is a binary variable associated with observation i. In inequalities (3) and (4), $c_i = \max\{\mathbf{b}_{k'}\mathbf{x}_i \mid k' \neq k\}$ is sufficient for any $i \in I_k$. If these inequalities still hold with $d_i = 0$, observation $i \in I_k$ is correctly classified into class k. Otherwise, if $d_i = 1$ has to be held for these inequalities to hold, then observation $i \in I_k$ is incorrectly classified into a class other than k.

Similar to the LP formulation in Sun [2002b], the following MIP formulation is proposed to estimate the parameters, *i.e.*, the elements of $\boldsymbol{\beta}_k$ for all $k \in K$, for multiple-class discriminant analysis

$$min \quad \sum_{i \in I} d_i \tag{5}$$

$$s.t. \quad \mathbf{b}_k\mathbf{x}_i - c_i + Md_i \geq \varepsilon \qquad \text{for } i \in I_k \text{ and } k \in K \tag{6}$$

$$\mathbf{b}_k\mathbf{x}_i - c_i \leq 0 \qquad \text{for } i \notin I_k \text{ and } k \in K \tag{7}$$

$$\mathbf{b}_k \text{ unrestricted} \qquad \text{for } k \in K \tag{8}$$

$$c_i \text{ unrestricted} \qquad \text{for } i \in I \tag{9}$$

$$d_i = 0 \text{ or } 1 \qquad \text{for } i \in I. \tag{10}$$

Totally m binary variables are introduced into this MIP model, one for each observation in the sample. In (6), $\varepsilon > 0$ is a constant that is sufficiently small such that $M \geq \varepsilon$. In the implementation, it is sufficient to have $\varepsilon = 1$. Intuitively, ε creates a classification gap for each $i \in I$. In addition, ε plays a normalization role to eliminate trivial solutions.

According to (6) and (7), $(\mathbf{b}_k\mathbf{x}_i + Md_i) - \max\{\mathbf{b}_{k'}\mathbf{x}_i \mid k' \neq k\} \geq \varepsilon$ must hold for each $i \in I_k$. If $d_i = 0$ in the final solution, then $\mathbf{b}_k\mathbf{x}_i - \max\{\mathbf{b}_{k'}\mathbf{x}_i \mid k' \neq k\} \geq \varepsilon$ and observation $i \in I_k$ is correctly classified into class k with a clear margin greater than or equal to ε. Otherwise, if $d_i = 1$ in the final solution, $\mathbf{b}_k\mathbf{x}_i - \max\{\mathbf{b}_{k'}\mathbf{x}_i \mid k' \neq k\} \geq \varepsilon$ does not hold. In this case, observation $i \in I_k$ cannot be correctly classified into class k. Hence, the value of d_i indicates if the observation $i \in I_k$ is correctly classified. The objective function (5) represents the total number of misclassified observations, or the L_0 -norm. Because the objective function is minimized, as many as possible d_i are set to 0 in the optimal solution of the MIP model.

For each observation $i \in I$, one constraint in (6) and $p-1$ constraints in (7) are in the model. Altogether, there are m constraints in (6), one for each observation $i \in I$, and $m(p-1)$ constraints in (7), $p-1$ for each observation $i \in I$. In addition to the binary variables d_i, the estimated parameters, *i.e.*, the elements of $\mathbf{b}_k$, and the cutoff points c_i are the continuous variables of the model. Altogether the model has mp constraints, m binary variables, and $(n+1)p + m$ continuous variables.

The MIP formulation in (5)-(10) is different from that in Gehrlein [1986]. In the MIP model in Gehrlein [1986], each observation $i \in I_k$ is associated with $p-1$ constraints, one for each $k' \neq k$. For an observation $i \in I_k$, each constraint is of the form

$$\mathbf{b}_k \mathbf{x}_i - \mathbf{b}_{k'} \mathbf{x}_i + M d_{ik'} \geq \varepsilon \qquad \text{for } i \in I_k,\ k \in K,\ k' \in K \text{ and } k' \neq k. \tag{11}$$

The constraint in (11) is the difference between the constraint in (6) for the observation and a constraint in (7) for the same observation but for a $k' \neq k$. As a result, $p-1$ binary variables of the form $d_{ik'}$ are used for each observation $i \in I_k$, one for each $k' \neq k$, in the model in Gehrlein [1986]. Totally $m(p-1)$ binary variables are in the MIP model of Gehrlein [1986] but only m are in the model in (5)-(10). Each x_{ij} appears $2(p-1)$ times in the MIP model of Gehrlein [1986] but appears only p times in the MIP model in (5)-(10). Using the property of order preserving under shifting and positive rescaling sated in the next section, each x_{ij} appears in the MIP model only $p-1$ times. Therefore, the MIP model in (5)-(10) proposed in this study is a much sparser model than previous MIP formulations.

When $p = 2$, each observation $i \in I$ is associated with one constraint in (6) and one constraint in (7). The inequality $\mathbf{b}_2 \mathbf{x}_i - \mathbf{b}_1 \mathbf{x}_i - M d_i \leq -\varepsilon$ is obtained after subtracting the constraint in (6) from the constraint in (7) for each $i \in I_1$. Similarly, $\mathbf{b}_2 \mathbf{x}_i - \mathbf{b}_1 \mathbf{x}_i + M d_i \geq \varepsilon$ is obtained after subtracting the constraint in (7) from the constraint in (6) for each $i \in I_2$. With $\mathbf{b} = \mathbf{b}_2 - \mathbf{b}_1$, the MIP model in (5)-(10) for $p = 2$ can be written as,

$$\textit{min} \quad \sum_{i \in I} d_i \tag{12}$$

$$\textit{s.t.} \quad \mathbf{b}\mathbf{x}_i - M d_i \leq -\varepsilon \qquad \text{for } i \in I_1 \tag{13}$$

$$\mathbf{b}\mathbf{x}_i + M d_i \geq \varepsilon \qquad \text{for } i \in I_2 \tag{14}$$

$$\mathbf{b} \text{ unrestricted} \tag{15}$$

$$d_i = 0 \text{ or } 1 \qquad \text{for } i \in I. \tag{16}$$

The MIP model in (12)-(16) is similar to that in Stam and Joachimsthaler [1990]. The difference is that $\varepsilon = 0$ is used in Stam and Joachimsthaler [1990]. In this sense, the MIP model in (5)-(10) may be considered as a generalization of the one proposed by Stam and Joachimsthaler [1990] for two-class classification.

When $p = 2$, the model in (12)-(16) constructs an equation $\mathbf{b}\mathbf{x} = 0$ that represents a hyperplane separating the two classes in the prediction variable space with one class on each side of the hyperplane. When $p > 2$, the MIP model in (5)-(10) constructs p discriminant functions. The equation representing the hyperplane separating two classes is obtained by equating the pair of discriminant functions, or subtracting one from the other, representing the two classes.

In a similar manner, other formulations for the two-class problems can be easily generalized to the multiple-class problems. More formulations for the two-class problems are summarized in Erenguc and Koehler [1990], Joachimsthaler and Stam [1990], and Stam [1997].

MIP formulations for discriminant problems are NP-complete [Chen and Mangasarian, 1996] and, therefore, are computationally more demanding than LP formulations, especially when the classes have substantial

overlaps and the misclassification rate is high. This limitation may be overcome by developing heuristic solution methods to tackle such MIP formulations. The MIP model may have many alternative optimal solutions, especially when the classes in the sample are completely separable. However, when the classes in the sample are not completely separable, the discriminant functions constructed with an optimal solution of this model always yield the highest classification rate of the observations in the sample among all models of similar structure.

4. Some Properties of the MIP Model

The MIP model in (5)-(10) possesses some properties that the LP model [Sun, 2002b] has and some others that the LP model does not have. These properties are around the four difficulties encountered in earlier MP models for discriminant and classification analysis. In this section, some properties that the MIP model in (5)-(10) has but the LP model does not have are presented first and the properties that both the MIP model and the LP model have are discussed briefly.

The first issue to consider is if the MIP model in (5)-(10) has an optimal solution or even has any feasible solution. Theorem 1 in the following guarantees that an optimal solution can always be found.

Theorem 1: An optimal solution to the MIP model in (5)-(10) always exists if $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$.

Proof: For an optimal solution to exist, the objective function must be bounded and a feasible solution must exist. From (10), the objective function (5) is bounded from below by 0. Because of $M \geq \varepsilon$, the trivial solution with $\mathbf{b}_k = \mathbf{0}$ for all $k \in K$, $c_i = 0$ for all $i \in I$ and $d_i = 1$ for all $i \in I$ is a feasible solution. □

The objective function is also bound above by m. Hence, the largest possible value that the objective function may have is m. Therefore, a solution with $d_i = 1$ for all $i \in I$ is the possible worst feasible solution. Only an improper solution has such an objective function value. With an improper solution, $\mathbf{b}_k = \bar{\mathbf{b}}$ for all $k \in K$ and the discriminant functions cannot definitely classify any observation into any class. Although Theorem 1 insures an optimal solution always exists, the solution is useless if it is an improper solution. The second issue of concern is whether the optimal solution is an improper solution. Theorem 2 in the following insures that the MIP model in (5)-(10) always has proper solutions. The MIP model can always find proper solutions because proper solutions are always better than improper solutions, *i.e.*, always have lower values for the objective function.

Theorem 2: If $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$, the MIP model in (5)-(10) always has proper solutions.

Proof: Assume an improper solution with $\mathbf{b}_k = \bar{\mathbf{b}}$ for all $k \in K$ is optimal. The constraints in (7) are satisfied with $c_i = \bar{\mathbf{b}}\mathbf{x}_i$ for all $i \in I$. Hence, with $M \geq \varepsilon$, the constraints in (6) are satisfied only with $d_i = 1$ for all $i \in I$. As a

result, the objective function (5) has a value $\sum_{i \in I} d_i = m$. If all the observations are classified into a single class $k \in K$ such that $m_k = \max \{m_{k'} \mid k' \in K\}$, the objective function (5) has a value $\sum_{i \in I} d_i = m - m_k < m$. For any observation i to be classified into a class k, $\mathbf{b}_k \mathbf{x}_i > \mathbf{b}_{k'} \mathbf{x}_i$, *i.e.*, $\mathbf{b}_k \neq \mathbf{b}_{k'}$, must hold for all $k' \neq k$. This contradicts the assumption that the improper solution with $\mathbf{b}_k = \overline{\mathbf{b}}$ for all $k \in K$ is optimal.□

Because the trivial solution is a special case of improper solutions, with Theorem 2, the MIP model in (5)-(10) never generates the trivial solution. Many LP models for two-class classification suffer from trivial solutions [Freed and Glover, 1986b; Koehler, 1989a, 1989b, 1990, 1991; Markowski and Markowski, 1985]. For multiple-class discriminant and classification analysis, the LP approaches [Bennett and Mangasarian, 1994; Gochet, Stam, Srinivasan and Chen, 1997; Sun, 2002b] generate improper solutions $\mathbf{b}_k = \overline{\mathbf{b}}$ for all $k \in K$ under the special condition that all classes in the sample have the same class centroid and equal sample size. The centroid of a class $k \in K$ is defined to be $\overline{\mathbf{x}}_k = \sum_{i \in I_k} \mathbf{x}_i$. Hence, these LP models generate improper solutions when $\overline{\mathbf{x}}_1 = \overline{\mathbf{x}}_2 = ... = \overline{\mathbf{x}}_p$ and $m_1 = m_2 = ... = m_p$. However, even under this special condition, the MIP model in (5)-(10) always generates a proper solution with a set of discriminant functions that can separate the classes in the sample as much as possible.

The proof of Theorem 2 assumes that the discriminant functions classify all the observations in the sample into the class with the largest number of observations. In fact, better solutions can always be found if the different classes have different observations in the sample. Assume all observations are classified into class k. If there is an observation $i \in I_{k'}$ with $k' \neq k$ that appears only in class k', then reassigning this observation to class k' will reduce the value of the objective function (5) by 1. Under the extreme condition that all the classes have exactly the same observations, the model still generates a proper solution but the discriminant functions can only classify all the observations into a single class. Under this extreme condition, the observations should not be differentiated.

With discriminant functions constructed with LP models, some observations in the sample fall into the classification gap. For such an observation $i \in I_k$, both $\mathbf{b}_k \mathbf{x}_i > c_i$ and $\mathbf{b}_k \mathbf{x}_i < c_i + \varepsilon$ hold. Such observations are correctly classified but without a clear margin ε. Theorem 3 in the following shows that this situation does not exist with discriminant functions constructed with the MIP model in (5)-(10). Hence, an observation may be either incorrectly classified or correctly classified with a clear margin ε. Therefore, for discrimination purpose, the optimal solution of the MIP model is at least as good as the optimal solution of the LP model in terms of the number of misclassified observations in the sample. Therefore, the MIP model may be used to improve the solution obtained with the LP model.

Theorem 3: If $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$, then the condition in (17) in the following does not hold for any $i \in I_k$ in an optimal solution,

$$0 < \mathbf{b}_k \mathbf{x}_i - c_i < \varepsilon . \tag{17}$$

Proof: Any observation $i \in I_k$ for which (17) holds must have $d_i = 1$ to satisfy the constraint in (6). Assume (17) holds in an optimal solution for at least one $i \in I_k$, therefore, $d_i = 1$. By (17), $\phi = \varepsilon/(\mathbf{b}_k \mathbf{x}_i - c_i) > 1$. Let $\hat{\mathbf{b}}_k = \alpha \mathbf{b}_k$ for all $k \in K$, where $\alpha \geq \phi > 1$. Then $\hat{g}_k(\mathbf{x}) = \hat{\mathbf{b}}_k \mathbf{x}$ for all $k \in K$ preserve the orders of $g_k(\mathbf{x})$ [Sun, 200b] and, therefore, $\hat{g}_k(\mathbf{x})$ for $k \in K$ can also be used as discriminant functions. Because $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small with $M \geq \varepsilon$, replacing $\mathbf{b}_k$ with $\hat{\mathbf{b}}_k = \alpha \mathbf{b}_k$ for each $k \in K$ and replacing c_i with $\hat{c}_i = \alpha c_i$ for all $i \in I$ in (6) and (7) will not violate the feasibility of the MIP model in (5)-(10). However, $\hat{\mathbf{b}}_k \mathbf{x}_i - \hat{c}_i = \alpha(\mathbf{b}_k \mathbf{x}_i - c_i) \geq \phi(\mathbf{b}_k \mathbf{x}_i - c_i) = \varepsilon$, which implies $d_i = 0$ is feasible. This contradicts the assumption that the solution is optimal. □

The MIP model in (5)-(10) also has other properties that the LP model [Sun, 2002b] has. One property is stability or invariability under linear data transformation. Linear data transformation is to transform each x_{ij} to y_{ij} using $y_{ij} = \alpha_j x_{ij} + \gamma_j$ for each $i \in I$ and $j \in J$. Both α_j and γ_j are scalars with $\alpha_j \neq 0$ and may be different for different $j \in J$. The transformed data y_{ij} instead of the original data x_{ij} are then used to set up the MIP model and to construct the discriminant functions because the transformed data may be easier to analyze. The MIP model in (5)-(10) is invariant under linear data transformation as long as $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$ and $M \geq |\alpha| \varepsilon$. After transformation, the estimated parameters in the discriminant functions become $\hat{\mathbf{b}}_k = (b_{k0} - \sum_{j \in J} (b_{kj} / \alpha_j)\gamma_j, b_{k1} / \alpha_1, ..., b_{kn} / \alpha_n)$ for all $k \in K$. Then $\hat{\mathbf{b}}_k \mathbf{y}$ for $k \in K$ are the discriminant functions using the transformed data. Being stable or invariant, $\hat{\mathbf{b}}_k \mathbf{y}$ for $k \in K$ preserve the order of $\mathbf{b}_k \mathbf{x}$, *i.e.*, $\hat{\mathbf{b}}_k \mathbf{y} \geq \hat{\mathbf{b}}_{k'} \mathbf{y}$ for any $k' \neq k$ if and only if $\mathbf{b}_k \mathbf{x} \geq \mathbf{b}_{k'} \mathbf{x}$ for the same $k' \neq k$. With this property, the classification results will always be the same whether the original or the transformed data are used. Stability or invariability under linear data transformation is an important desirable property because linear data transformation is a common technique in data preprocessing. Some LP models proposed for two-class classification do not have this property [Markowski and Markowski, 1985]. Being not invariant means the resulting classification functions using the transformed data may give different classification results from those using the original data.

Another property is the freedom in selecting a value for ε in (6). Varying the value of ε in (6) only rescales the coefficients in the resulting discriminant functions $\mathbf{b}_k$ for $k \in K$ and the cutoff values c_i for all $i \in I$ but does not affect the classification results of the resulting discriminant functions as long as $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$. Suppose $\varepsilon > 0$ in (6) is replaced with $\hat{\varepsilon} = \alpha\varepsilon$ for any $\alpha > 0$, hence $\hat{\varepsilon} > 0$, and M in (6) is changed accordingly to keep $M \geq \hat{\varepsilon}$. A new MIP model is formulated with $\hat{\varepsilon}$. Then $\hat{\mathbf{b}}_k = \alpha \mathbf{b}_k$ for $k \in K$ and $\hat{c}_i = \alpha c_i$ for $i \in I$ are optimal in the new MIP model if and only if $\mathbf{b}_k$ for $k \in K$ and c_i for $i \in I$ are optimal in the original MIP model. The binary variables y_i for $i \in I$ will be the same in both MIP models. Consequently, $\hat{\mathbf{b}}_k \mathbf{x}$ for $k \in K$ preserve the order of $\mathbf{b}_k \mathbf{x}$ for $k \in K$, *i.e.*, $\hat{\mathbf{b}}_k \mathbf{x} \geq \hat{\mathbf{b}}_{k'} \mathbf{x}$ for any $k' \neq k$ if and

only if $\mathbf{b}_k\mathbf{x} \geq \mathbf{b}_{k'}\mathbf{x}$ for the same $k' \neq k$. As a result, $\hat{\mathbf{b}}_k\mathbf{x}$ and $\mathbf{b}_k\mathbf{x}$ for $k \in K$ give the same classification results. With this property, users can use any convenient value for ε. The value of ε determines the magnitude of the components of $\mathbf{b}_k$ for $k \in K$ and the magnitude of c_i for $i \in I$.

Sun [2002b] studied the order preserving property of the discriminant functions for the LP models under shifting and positive rescaling. Shifting refers to adding to each discriminant function a constant and positive rescaling refers to multiplying each discriminant function by a positive constant. Discriminant functions constructed with statistical methods and LP models all have this property. If a function of the form $\mathbf{bx}$ is added to each discriminant function $g_k(\mathbf{x})$ to obtain $\hat{g}_k(\mathbf{x}) = (\mathbf{b}_k + \mathbf{b})\mathbf{x}$ for each $k \in K$, the functions $\hat{g}_k(\mathbf{x})$ and $g_k(\mathbf{x})$ have the same classification results. If each discriminant function $g_k(\mathbf{x})$ is multiplied by the same constant $\alpha > 0$ to obtain $\hat{g}_k(\mathbf{x}) = \alpha\mathbf{bx}$, the functions $\hat{g}_k(\mathbf{x})$ and $g_k(\mathbf{x})$ also have the same classification results. Sun [2002b] showed that the LP model can be simplified by using this property. The MIP model in (5)-(10) can also be simplified by using this property.

Choose a $k' \in K$ such that $m_{k'} = min\{m_k \mid k \in K\}$ and add $-\mathbf{b}_{k'}\mathbf{x}$ to $\mathbf{b}_k\mathbf{x}$ to obtain $\hat{\mathbf{b}}_k\mathbf{x} = (\mathbf{b}_k - \mathbf{b}_{k'})\mathbf{x}$ for all $k \in K$. Hence, $\hat{\mathbf{b}}_{k'} = \mathbf{0}$. Using $\hat{\mathbf{b}}_k = \mathbf{b}_k - \mathbf{b}_{k'}$ instead of $\mathbf{b}_k$ in the MIP model, the constraint in (6) for each $i \in I_{k'}$ becomes $-c_i + Md_i \geq \varepsilon$. Let $c_i' = -c_i$ for all $i \in I_{k'}$ in the MIP model. Then $c_i' \geq \varepsilon$ if $d_i = 0$ and $c_i' \geq \varepsilon - M$ if $d_i = 1$. By setting $c_i' \geq 0$, the MIP model will not lose generality because $\varepsilon - M$ can be absorbed by b_{k0} for $k \in K$. The constraint in (7) for each $i \notin I_{k'}$ and for the chosen k' becomes $c_i \geq 0$. As a result, the cutoff values, *i.e.*, the c_i' for $i \in I_{k'}$ and the c_i for $i \notin I_{k'}$, become nonnegative from unrestricted and all the constraints in (9) can be replaced with nonnegativity constraints. The number of constraints in (7) is reduced by $m - m_{k'}$ and $m_{k'}$ constraints in (6) become the form of $c_i' + Md_i \geq \varepsilon$. Therefore, the MIP model is simplified. By using this property, the number of constraints in the MIP model (5)-(10) is reduced from mp to $m(p-1) + m_{k'}$ and the number of continuous variables is reduced from $p(n+1) + m$ to $(p-1)(n+1) + m$. Choosing k' such that $m_{k'} = min\{m_k \mid 1 \leq k \leq p\}$ to set $\hat{\mathbf{b}}_{k'} = \mathbf{0}$ will take full advantage of this property.

The condition of the properties of the MIP model in (5)-(10) is that $M > 0$ is sufficiently large and $\varepsilon > 0$ is sufficiently small such that $M \geq \varepsilon$. No guidelines of determining M were provided for two class models published in the literature. The following is a rough guideline for determining M in (6). Solve the LP model in Sun [2002b] with the same ε. Suppose e_i^* for each $i \in I$ are the values of the deviation variables in an optimal solution of the LP model. Then the value of M can be determined by $M \geq M^* = max\{e_i^* \mid i \in I\}$. For practical problems, the LP model may have alternative optimal solutions. Each optimal solution may have a different M^* and the optimal solution of the MIP model may need a M that is larger than M^*. Because the CPU time needed to solve the LP model can be negligible as compared to that to solve the MIP model, this step does not cause extra computation burden. Furthermore, an optimal solution of the LP model provides an initial incumbent for the solution of the MIP model.

5. Examples

Some computational results of the MIP models in (5)-(10) are reported in this section. The main purpose of these examples is to show the ways that the model work, rather than to show the performance of this approach relative to other discriminant methods.

The software package CPLEX®[1] was used to solve the MIP problems. All computations were conducted on a SUN Enterprise 3000 computer running the UNIX operating system. Computer programs were written to convert the data sets in spreadsheet format to the MPS format that CPLEX® can read.

For the Iris Data Set and the Personnel Data Set, both in-sample results and validation results are reported. The in-sample results are the results obtained when all the observations in the sample are used to construct the discriminant functions and the resulting discriminant functions are used to classify each of the observations in the sample. The validation results were obtained using the leave one out validation procedure. With this validation procedure, one observation $i \in I$ is left out in turn and the other $m-1$ observations are used to construct the discriminant functions. The resulting discriminant functions are then used to classify the observation that is left out. This process is repeated m times, once for each $i \in I$.

5.1 A Structured Example

The purpose of this example is to demonstrate how to set up the MIP model in (5)-(10). The example has $m = 9$ observations and $n = 2$ characteristics as shown in Table 1. The 9 observations are divided equally into $p = 3$ classes. The observations in the 3 classes are not completely separable. The objective function of this example problem is

$$min \quad d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7 + d_8 + d_9 \tag{18}$$

With $M = 10$ and $\varepsilon = 1$ and with all coefficients in the first discriminant function set to 0, the constraints in (6) and (7) are listed in Table 2. The constraints in (6) are in the diagonal cells and those in (7) are in the off diagonal cells of the table. In addition, b_{kj} are unrestricted for all $j = 0,1,2$ and $k = 1,2,3$, $c_i \geq 0$ and $e_i \geq 0$ for $i = 1,...,9$. For each observation $i \in I_1$, c_i instead of c_i' is used to simplify the notation.

The discriminant functions obtained are, respectively, $g_1(\mathbf{x}) = 0$, $g_2(\mathbf{x}) = 9.5 - 2.09x_1 - 1.09x_2$ and $g_3(\mathbf{x}) = -24.56 - 9.02x_1 + 15.70x_2$. With this set of discriminant functions, observation 1 is misclassified into class 2 and observation 6 is misclassified into class 1, and all other observations are correctly classified. The equations representing the hyperplanes separating the three classes obtained from this set of discriminant functions are $2.09x_1 + 1.09x_2 = 9.5$ between classes 1 and 2, $-9.02x_1 + 15.70x_2 = 24.56$ between classes 1 and 3, and $-6.93x_1 + 16.80x_2 = 34.06$ between classes 2 and 3.

---

[1] CPLEX Optimization, Inc., *Using the CPLEX® Callable Library Including Using the CPLEX® Base System with CPLEX® Barrier and Mixed Integer Solver Options*, 1989-1995.

5.2 The Wine Recognition Data Set

This data set was originally used by Aeberhard, Coomans and de Vel [1992]. The data are the results of a chemical analysis of wines grown in the same region in Italy but derived from $p = 3$ different cultivars. The analysis determined the quantities of $n = 13$ constituents, *i.e.*, prediction variables, found in each of the $p = 3$ types of wines. The data set itself with more information is available at the UCI Repository of Machine Learning Databases [Merz and Murphy, 1998]. Because the observations of the three classes in the sample are completely separable, the problem is computationally easy to solve. The MIP formulation in (5)-(10) for this example took CPLEX 0.15 seconds to solve.

The discriminant functions obtained are $g_1(\mathbf{x}) = -346.25x_2 + 1766.04x_7 + 2128.27x_8 - 492.58x_{10} - 0.70x_{13}$, $g_2(\mathbf{x}) = 1591.52 - 382.30x_2 + 1.44x_5 + 1766.04x_7 + 1681.39x_8 + 85.58x_9 - 653.51x_{10} - 37.56x_{12} - 1.85x_{13}$, and $g_3(\mathbf{x}) = 0$ for the three classes respectively. The classification results are presented in Table 3. This data set has been used by other researchers to test different discriminant methods, *e.g.*, by Bennett and Mangasarian [1994]. The results obtained with the MIP approach are comparable with those published in the literature.

5.3 The Iris Data Set

This example was originally used by Fisher [1936] and was used as a standard test data set by many authors, such as Kendall [1966], Gehrlein [1986] and Bennet and Mangasarian [1994] and as a standard example in many commercial software packages, such as NeuralWorks [NeuralWare, 1993] and BMDP [Jennrich and Sampson, 1983]. The data set contains 4 prediction variables and 150 observations on 3 species of iris plants, Iris setosa, Iris versicolor, and Iris virginica. The 4 prediction variables measure 4 plant characteristics, sepal length (x_1), sepal width (x_2), petal length (x_3) and petal width (x_4). The 150 observations are evenly divided into the 3 classes with $m_k = 50$ in each. The data set is available in Kendall [1966] and Gehrlein [1986]. The LP model for this problem took CPLEX less than 0.1 seconds to solve. Because the 3 classes are nearly completely separable, the MIP model for this problem is not difficult to solve. It took CPLEX a little over 1 second.

With a $M = 10$ and an $\varepsilon = 1$, the discriminant functions obtained with the MIP model in (5)-(10) are $g_1 = 0.00$ for Iris setosa, $g_2(\mathbf{x}) = 169.84 + 8.36x_1 - 103.37x_2 + 97.59x_4$ for Iris versicolor, and $g_3(\mathbf{x}) = -117.55x_2 + 39.64x_3 + 138.82x_4$ for Iris virginica. The coefficients in the discriminant functions were rounded to the second decimal digit and those in the first function were all set to 0. The classification results are given in Table 4. These results are comparable with those obtained with other methods as published in the literature.

With different values of M in (6), the same in-sample classification rate was obtained each time with 149 out of the 150 observations correctly classified. Sometimes one observation in Iris versicolor was incorrectly classified into Iris virginica and sometimes one observation in Iris virginica was incorrectly classified into Iris versicolor, but not both. The discriminant functions of both the single function and multiple function MIP models of Gehrlein [1986] classified 149 observations to their correct classes with one observation from Iris versicolor misclassified into Iris virginica. Gehrlein [1986] misclassified the same observation in Iris versicolor. With the

leave one out validation procedure, 143 out of 150 observations were correctly classified. Fisher's LDF resulted in 3 observations misclassified on this data set [Jennrich and Sampson, 1983].

It is not a surprise for the discriminant functions generated by the MIP models to achieve a better in-sample classification rate but a worse validation classification rate because the objectives of these MIP models are to minimize the number of misclassifications in the sample and the objective of the LP model is to minimize the sum of deviations from the cutting points of the misclassified observations.

5.4 The MBA Admission Data Set

Johnson and Wichern [1988] provided a MBA Admission Data Set as an example for multiple-class discriminant analysis. There are a total of 85 observations divided into 3 classes of applicants, "Admit" (Class 1), "Not admit" (Class 2) and "Borderline" (Class 3), of a business school. Two variables, undergraduate GPA (x_1) and GMAT score (x_2), are used to measure each observation. Among the 85 applicants in the sample, 31 are in the class "Admit", 28 are in the class "Not admit" and the other 26 are in the class "Borderline". The details of the data set are described and the data set itself is available in the book [Johnson and Wichern, 1988].

Because the observations in the three classes are nearly separable, the MIP model is easy to solve. It took CPLEX less than 0.1 second of CPU time to solve. With $M = 10$ and $\varepsilon = 1$ in the MIP model in (5)-(10), the discriminant functions obtained are $g_1(\mathbf{x}) = -303.8947 + 56.3158x_1 + 0.2632x_2$ for the class "admit", $g_2(\mathbf{x}) = 4441.7742 - 1199.5526x_1 - 2.5054x_2$ for the class "Not admit" and $g_3(\mathbf{x}) = 0$ for the class "Borderline" after the parameters are rounded to the fourth decimal digit and those in the third discriminant function are set to 0.

The classification results are summarized in Table 5. This data set has been used to test other MP models for discriminant analysis in different studies, such as in Loucopoulos [2001]. The results obtained by the MIP model in (5)-(10) are in line with those published in the literature.

5.4 Another Structured Example

The data of this example are presented in Table 6. The samples of this data set have the same class centroid and equal sample size, *i.e.*, $\bar{\mathbf{x}}_1 = \bar{\mathbf{x}}_2 = \bar{\mathbf{x}}_3 = (7.126, 17.389, 12.960)$ and $m_1 = m_2 = m_3 = 10$. Therefore, this data set meets the conditions for improper solutions of LP models [Bennett and Mangasarian, 1994; Gochet, Stam, Srinivasan and Chen, 1997; Sun, 2002b]. As a result, these LP models cannot generate any meaningful discriminant functions for this data set. Fisher's LDF also generated an improper solution for this data set.

With $M = 10$ and $\varepsilon = 1$, the MIP model generated the discriminant functions $g_1(\mathbf{x}) = 0$ for class 1, $g_2(\mathbf{x}) = 125.45 - 28.68x_1 + 14.91x_2 - 14.58x_3$ for class 2, and $g_3(\mathbf{x}) = -2641.33 - 49.71x_1 + 139.11x_2 + 35.68x_3$ for class 3, respectively. The coefficients are all rounded to the second decimal digit and those in the first discriminant function are all set to 0. The in-sample classification results are presented in Table 7. The in-sample classification rate is 63.33%. Given the small sample size, no meaningful validation results were obtained for this example.

6. Conclusions

A MIP formulation for multiple-class discriminant and classification analysis is proposed, that directly minimizes the number of misclassifications in the sample. The formulation is simple, easy to understand and easy to use. Properties of the model are discussed. The model does not suffer from any difficulties caused by pathologies of some of the MP formulations for two-class classification analysis. With this MIP approach, practitioners have one more technique in analyzing their discriminant problems.

In general, LP models are preferred to MIP models because LP models are much easier to solve and the resulting discriminant functions may have better generalization capabilities for new observation classification. Under some conditions the MIP approach may be preferable. For example, the MIP approach may be preferred to the LP approach if the purpose of the application is discrimination rather than classification. Although MIP models are generally much more difficult to solve, they can be solved within reasonable computation time under certain conditions. For example, the MIP models are not difficult to solve when the sample sizes are small and when the observations of the different classes in the sample are completely or nearly completely separable.

One direction of future research in this area is computational experiments to test the performance of the MIP model proposed in this study relative to other approaches under different data conditions. It is also necessary to determine the effect of the relative values of M and ε in (6) on the computational complexity of the MIP model through computational experiments. One direction is to address the issue of variable selection when observations on a large number of variables are available [Sun and Xiong, 2002a, 2002b]. Using the MIP formulation, one more set of binary variables will be involved in the variable selection model and, therefore, the model will demand more computation time. Another direction is to develop heuristic methods to solve the MIP models, possibly with variable selection capability, especially for applications with large data sets. With effective heuristics, the disadvantage of demanding too much computation time is at least partially overcome. One more direction is software implementations. The MP approaches will be much easier for the practitioners to use if user friendly software is available, possibly with variable selection features and with heuristic procedures to solve MIP models. If the observations in the sample are completely or nearly completely separable, the MIP model may not take much more time than LP models.

Acknowledgement

This research was supported in part by a grant from the College of Business at University of Texas at San Antonio.

References

Aeberhard, S., D. Coomans and O. de Vel, "Comparison of Classifiers in High Dimensional Settings," Technical Report No. 92-02, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland, 1992.

Abad, P. L. and W. J. Banks, "New LP Based Heuristics for the Classification Problem," *European Journal of Operational Research*, Vol. 67, No. 1, pp. 88-100, 1993.

Alman, E. I., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, Vol. 23, pp. 589-609, 1968.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 96, pp. 6745-6750, 1999.

Bajaier, S. M. and A. V. Hill, "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem," *Decision Sciences*, Vol. 13, pp. 604-618, 1982.

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

Bennett, K. P. and O. L. Mangasarian, "Multicategory Discrimination via Linear Programming," *Optimization Methods and Software*, Vol. 3, pp. 27-39, 1994.

Breiman, L. J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

Cavalier, T. M., J. P. Ignizio and A. L. Soyster, "Discriminant Analysis via Mathematical Programming on Certain Problems and Their Causes," *Computers & Operations Research*, Vol. 16, No. 4, pp. 353-362, 1989.

Chen, C. and O. L. Mangasarian, "Hybrid Misclassification Minimization," *Advances in Computational Mathematics*, Vol. 5, No. 2, pp. 127-136, 1996.

Choo, E. -U. and W. C. Wedley, "Optimal Criteria Weights in Repetitive Multicriteria Decision-Making," *Journal of the Operational Research Society*, Vol. 36, pp. 983-992, 1985.

Conwey, D. G., V. A. Cabot and M. A. Venkataramanan, "A Genetic Algorithm for Discriminant Analysis," *Annals of Operations Research*, Vol. 78, pp. 71-82, 1998.

Duarte Silva, A. P. and A. Stam, "Second-Order Mathematical Programming formulations for Discriminant Analysis," *European Journal of Operational Research*, Vol. 74, No. 1, pp. 4-22, 1994.

Dudoit, S., J. Fridlyand and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of American Statistical Association*, Vol. 97, No. 457, pp. 77-87, 2002.

Dutka, A. *AMA Handbook of Customer Satisfaction: A Guide to Research, Planning and Implementation*, NTC Publishing Class, Illinois, 1995.

Erenguc, S. S. and G. J. Koehler, "Survey of Mathematical Programming Models and Experimental Results for Linear Discriminant Analysis," *Management and Decision Economics*, Vo. 11, pp. 215-225, 1990.

Fisher, R. A., "The Use of Multiple Measurements in Taxonomy Problems," *Annals of Eugenics*, Vol. 7, pp. 179-188, 1936.

Fraughnaugh, K. J. Ryan, H. Zullo and L. A. Cox, "Heuristics for Efficient Classification," *Annals of Operations Research*, Vol. 78, pp. 189-200, 1998.

Freed, N. and F. Glover, "A Linear Programming Approach to the Discriminant Problem," *Decision Sciences*, Vol. 12, No. 1, pp. 68-74, 1981a.

Freed, N. and F. Glover, "Simple but Powerful Goal Programming Formulations for the Discriminant Problem," *European Journal of Operational Research*, Vol. 7, No. 1, pp. 44-60, 1981b.

Freed, N. and F. Glover, "Evaluating Alternative Linear Programming Models to Solve the Two-Class Discriminant Problem," *Decision Sciences*, Vol. 17, No. 2, pp. 151-162, 1986a.

Freed, N. and F. Glover, "Resolving Certain Difficulties and Improving the Classification Power of LP Discriminant Analysis Formulations," *Decision Sciences*, Vol. 17, No.4, pp. 589-595, 1986b.

Gehrlein, W. V., "General Mathematical Programming Formulations for the Statistical Classification Problem," *Operations Research Letters*, Vol. 5, No. 6, pp. 299-304, 1986.

Gehrlein, W. V. and B. J. Wagner, "Nontraditional Approaches to the Statistical Classification and Regression Problems," Annals of Operations Research, Vol. 74, 1997.

Glen, J. J., "Integer Programming Methods for Normalization and Variable Selection in Mathematical Programming Discriminant Analysis Models," *The Journal of the Operational Research Society*, Vol. 50, No. 10, pp. 1043-1053, 1999.

Glen, J. J., "Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach," *The Journal of the Operational Research Society*, Vol. 52, No. 3, pp. 328-339, 2001.

Glover, F., "Improved Linear Programming Models for Discriminant Analysis," *Decision Sciences*, Vol. 21, No. 4, pp. 771-785, 1990.

Glover, F., S. Keene, and B. Duea, "A New Class of Models for the Discriminant Problem," *Decision Sciences*, Vol. 19, No. 2, pp. 269-280, 1988.

Gochet, W., A., Stam, V. Srinivasan and S. Chen, "Multiclass Discriminant Analysis Using Linear Programming," *Operations Research*, Vol. 45, No.2, pp.213-225, 1997.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, Vol. 286, pp. 531-537, 1999.

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

Happer, P. R., "Patient Classification and the Port Package," *Safety Science Monitor*, Vol. 9, No. 1, 2005.

Huberty, C. J., *Applied Discriminant Analysis*, Wiley, New York, 1994.

Jennrich, R. and P. Sampson, "Stepwise Discriminant Analysis," in W. J. Dixon (ed.), *BMDP Statistical Software*, University of Califonial Press, Berkeley, CA, pp. 519-525, 1983.

Joachimsthaler, E. A. and A. Stam, "Four Approaches to the Classification Problem in Discriminant Analysis: An Experimental Study," *Decision Sciences*, Vol. 19, No.2, pp. 322-333, 1988.

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

Johnson, D. E., *Applied Multivariate Methods for Data Analysis, Duxbury Press*, Pacific Grove, CA, 1998.

Johnson, R. A. and D. W. Wichern, *Applied multivariate statistical analysis*, Second Edition, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

Kendall, M. G., "Discrimination and Classification," in P. R. Krishnaiah (ed.), *Multivariate Analysis*, Academic Press, New York, New York, pp. 165-185, 1966.

Koehler, G. J., "Characterizations of Unacceptable Solutions in LP Discriminant Analysis," *Decision Sciences*, Vol. 20, No. 2, pp. 239-257, 1989a.

Koehler, G. J., "Unacceptable Solutions and the Hybrid Discriminant Model," *Decision Sciences*, Vol. 20, No. 4, pp. 844-848, 1989b.

Koehler, G. J., "Considerations for Mathematical Programming Models in Discriminant Analysis," *Managerial and Decision Economics*, Vol. 11, pp. 227-234, 1990.

Koehler, G. J., "Improper Linear Discriminant Classifiers," *European Journal of Operational Research*, Vol. 50, pp. 188-198, 1991.

Lam, K. F., E. U. Choo and J. W. Moy, "Minimizing Deviations from the Class Mean: A New Linear Programming Approach for the Two-Class Classification Problem," *European Journal of Operational Research*, Vol. 88, pp. 358-367, 1998.

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

Lam, K. F. and J. W. Moy, "An Experimental Comparison of Some Recently Developed Linear Programming Approaches to the Discriminant Problem," *Computers & Operations Research*, Vol. 24, No. 7, pp. 593-599, 1997.

Lam, K. F. and J. W. Moy, "Combining Discriminant Methods in Solving Classification Problems in Two-Class Discriminant Analysis," *European Journal of Operational Research*, Vol. 138, No. 2, pp. 294-301, 2002.

Loucopoulos, C., "Three-Class Classification with Unequal Misclassification Costs: A Mathematical Programming Approach," *Omega*, Vol. 29, No.3, pp. 291-297, 2001.

Loucopoulos, C. and R. Pavur, "Computational Characteristics of a New Mathematical Programming Model for the Three-Class Discriminant Problem," *Computers & Operations Research*, Vol. 24, No. 2, pp. 179-191, 1997a.

Loucopoulos, C. and R. Pavur, "Experimental Evaluation of the Classificatory Performance of Mathematical Programming Approaches to the Three-Class Discriminant Problem: The Case of Small Samples," *Annals of Operations Research*, Vol. 74, pp. 191-209, 1997b.

Markowski, E. P. and C. A. Markowski, "Some Difficulties and Improvements in Applying Linear Programming Formulations to the Discriminant Problem," *Decision Sciences*, Vol. 16, No.3, pp. 237-247, 1985.

Markowski, C. A. and P. E. Markowski, "An Experimental Comparison of Several Approaches to the Discriminant Problem with Both Qualitative and Quantitative Variables," *European Journal of Operational Research*, Vol. 28, pp. 74-87, 1987.

Merz, C. J. and P.M. Murphy, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/mlearn /MLRepository.html., 1998.

Nakayama, H. and N. Kagaku, "Pattern Classification by Linear Goal Programming and Its Extensions," *Journal of Global Optimization*, Vol.12, No. 2, pp. 111-126, 1998.

Nath, R., W. M. Jackson and T. W. Jones, "A Comparison of the Classical and the Linear Programming Approaches to the Classification Problem in Discriminant Analysis," *Journal of the Statistical Computation and Simulation*, Vol. 41, pp. 73-93, 1992.

Nath, R. and T. W. Jones, "A Variable Selection Criterion in Linear Programming Approaches to Discriminant Analysis," *Decision Sciences*, Vol. 19, No. 3, pp. 554-563, 1988.

NeuralWare, Using NeuralWorks, A Tutorial for NeuralWorks Professional II/Plus and NeuralWorks Explorer, NeuralWare, Inc., Pittsburgh, Pennsylvania, 1993.

Östermark, R. and R. Höglund, "Addressing the Multiclass Discriminant Problem Using Multivariate Statistics and Mathematical Programming," *European Journal of Operational Research*, Vol. 108, No. 1, pp. 224-237, 1998.

Pavur, R., "Dimensionality Representation of Linear Discriminant Function Space for the Multiple-Class Problem: An MIP Approach," *Annals of Operations Research*, Vol. 74, pp. 37-50, 1997.

Pavur, R. and C. Loucopoulos, "Examining Optimal Criterion Weights in Mixed Integer Programming Approaches to Multiple-class Classification Problem," *Journal of the Operational Research Society*, Vol. 46, pp. 626-640, 1995.

Pavur, R. and C. Loucopoulos, "Evaluating the Effect of Gap Size in a Single Function Mathematical Programming Model for the Three-Class Classification Problem," *The Journal of the Operational Research Society*, Vol. 52, No. 8, pp. 896-904, 2001.

Ragsdale, C. T. and A. Stam, "Mathematical Programming Formulations for the Discriminant Problem: An Old Dog Does New Tricks," *Decision Sciences*, Vol. 22, No. 2, pp. 296-307, 1991.

Rossi, L., R. Slowinski and R. Susmanga, "Rough Set Approach to Evaluation of Stormwater Pollution," *International Journal of Environment and Pollution*, Vol. 12, No. 2-3, pp. 232-250, 1999.

Rubin, P. A., "Separation Failure in Linear Programming Discriminant Models," *Decision Sciences*, Vol. 20, pp. 519-535, 1989a.

Rubin, P. A., "Evaluating the Maximize Minimum Distance Formulation of the Linear Discriminant Problem," *European Journal of Operational Research*, Vol. 41, pp. 240-248, 1989b.

Rubin, P. A., "Heuristic Solution Procedures for a Mixed-Integer Programming Discriminant Model," *Managerial and Decision Economics*, Vol. 11, pp. 255-266, 1990a.

Rubin, P. A., "A Comparison of Linear Programming and Parametric Approaches to the Two-Class Discriminant Problem," *Decision Sciences*, Vol. 21, pp. 373-386, 1990b.

Rubin, P. A., "A Comment Regarding Polynomial Discriminant Analysis," *European Journal of Operational Research*, Vol. 72, No. 1, pp. 29-31, 1994.

Rulon, P. J., D. V. Tiedeman, M. M. Tatsuoka and C. R. Langmuir, *Multivariate Statistics for Personnel Classification*, Wiley, New York, New York, 1967.

SAS Institute, Inc., *SAS/STAT User's Guide*, Release 6.03 Edition, SAS Institute, Carey, NC, 1988.

Shankar, B. U., S. K. Meher, A. Ghosh and L. Bruzzone, "Remote Sensing Image Classification: A Neuro-fuzzy MCS Approach," *Lecture Notes in Computer Science*, Vol. 4338, pp. 128-139, 2007.

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

Srinivasan, V. and Y. H. Kim, "Credit Granting: A Comparative Analysis of Classification Procedures," *Journal of Finance*, Vol. 42, pp. 665-683, 1987.

Srinivasan, V. and A. Shocker, "Linear Programming Techniques for Multi-Dimentional Analysis of Preferences," *Psychometrica*, Vol. 38, pp. 337-369, 1973.

Stam, A. "Extensions of Mathematical Programming-Based Classification Rules: A Multicriteria Approach," *European Journal of Operational Research*, Vol. 48, No. 3, pp. 351-361, 1990.

Stam, A. "Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p -norm Methods," *Annals of Operations Research*, Vol. 74, pp. 1-36, 1997.

Stam, A., and E. A. Joachimsthaler, "A Comparison of a Robust Mixed-Integer Approach to Existing Methods for Establishing Classification Rules for the Discriminant Problem," *European Journal of Operational Research*, Vol. 46, No. 1, pp. 113-122, 1990.

Stam, A. and C. T. Ragsdale, "On the Classification Gap in MP-Based Approaches to the Discriminant Problem," *Naval Research Logistics*, Vol. 39, pp. 545-559, 1992.

Stam, A. and D. R. Ungar, "RANGU: A Micocomputer Package for Two-Class Mathematical Programming-Based Nonparametric Classification," *European Journal of Operational Research*, Vol. 86, No. 2, pp. 374-388, 1985.

Sun, M., "A Multiple Objective Programming Approach for Determining Faculty Salary Equity Adjustments," *European Journal of Operational Research*, Vol. 138, No. 2, pp. 302-319, 2002a.

Sun, M., "Linear Programming Approaches for Multiple-Class Discriminant and Classification Analysis," Working Paper, College of Business, The University of Texas at San Antonio, San Antonio, Texas, 78249, 2002b.

Sun, M. and P. G. McKeown, "Tabu Search Applied to the General Fixed Charge Problem," *Annals of Operations Research*, Vol. 41, pp. 405-420, 1993.

Sun, M. and M. Xiong, "A Mathematical Programming Approach for Gene Selection and Tissue Classification," Working Paper, College of Business, The University of Texas at San Antonio, San Antonio, Texas, 78249, 2002a.

Sun, M. and M. Xiong, "A Mathematical Programming Approach for Gene Selection and Tissue Classification with Multiple Classes," Working Paper, College of Business, The University of Texas at San Antonio, San Antonio, Texas, 78249, 2002b.

Shankar, B. U., S. K. Meher, A. Ghosh and L. Bruzzone, "Remote Sensing Image Classification: A Neuro-fuzzy MCS Approach," *Lecture Notes in Computer Science*, Vol. 4338, pp. 128-139, 2007.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, New York, New York: Springer.

Vapnik, V. N. (1998), *Statistical Learning Theory*, New York, New York: Wiley.

Walker, R. B., "Discriminant and Classification Analysis as an Aid to Employee Selection," *Accounting Review*, Vol. 49, pp. 514-523, 1974.

Wilson, J. M., "Integer Programming Formulation of Statistical Classification Problems," *Omega*, Vol. 24, No. 6, pp. 681-688, 1996.

Xiao, B., "Necessary and Sufficient Conditions of Unacceptable Solutions in LP Discriminant Analysis," *Decision Sciences*, Vol. 24, pp. 699-712, 1993.

Xiao, B., "Necessary and Sufficient Conditions of Unacceptable Solutions in NLP Discriminant Analysis," *European Journal of Operational Research*, Vol. 77, pp. 404-412, 1994.

Xiao, B., "Decision Power and Solutions of LP Discriminant Models: Rejoinder," *Decision Sciences*, Vol. 25, pp. 335-336, 1994.

Xiong, M., X. Fang and J. Zhao, "Biomarker Identification by Feature Wrappers", *Genome Research*, Vol. 11, No. 11, pp. 1878-1887, 2001.

Xiong, M., L. Jin, W. Li and E. Boerwinkle, "Tumor Classification Using Gene Expression Profiles," *Biotechniques*, Vol. 29, pp. 1264–1270, 2000.

Xiong, M., W. Li, J. Zhao, L. Jin and E. Boerwinkle, "Feature (Gene) Selection in Gene Expression-Based Tumor Classification," *Molecular Genetics and Metabolism*, Vol. 73, pp. 239–247, 2001.

Yanev, N. and S. Balev, "A Combinatorial Approach to the Classification Problem," *European Journal of Operational Research*, Vol. 115, No. 2, pp. 339-350, 1999.

Yin, Q. and P. Guo, "Multispectral Remote Sensing Image Classification with Multiple Features,"*Proceedings of International Conference on Machine Learning and Cybernetics*, Vol 1, No., 19-22, Pp. 360 – 365, 2007.

Zhang, H., C. -Y. Yu, B. Singer and M. Xiong, "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 98, No. 12, pp. 6730-6735, 2001.

Zopounidis, C., *Operational Tools in the Management of Financial Risks*, Kluwer Academic Publishers, Dordrecht, 1998.

Zopounidis C. and A. I. Dimitras, *Multicriteria Decision Aid Methods for the Prediction of Business Failure*, Kluwer Academic Publishers, Dordrecht, 1998.

Zopounidis, C. and M. Doumpos, "Multicriteria Classification and Sorting Methods: A Literature Review," *European Journal of Operational Research*, Vol. 138, No. 2, pp. 229-246, 2002.

Table 1. Data in Example 1

Class 1			Class 2			Class 3		
Observation	Variables		Observation	Variables		Observation	Variables	
	1	2		1	2		1	2
1	2.0	2.0	4	2.5	3.0	7	1.0	2.5
2	5.5	4.5	5	2.0	1.5	8	5.0	4.5
3	4.5	1.0	6	6.5	4.5	9	2.5	4.0

Class	Obs.	Discriminant Function		
		$k=1$	$k=2$	$k=3$
$k=1$	1	$c_1 + 10\,d_1 \geq 1.0$	$b_{20} + 2.0\,b_{21} + 2.0\,b_{22} + c_1 \leq 0.0$	$b_{30} + 2.0\,b_{31} + 2.0\,b_{32} + c_1 \leq 0.0$
	2	$c_2 + 10\,d_2 \geq 1.0$	$b_{20} + 5.5\,b_{21} + 4.5\,b_{22} + c_2 \leq 0.0$	$b_{30} + 5.5\,b_{31} + 4.5\,b_{32} + c_2 \leq 0.0$
	3	$c_3 + 10\,d_3 \geq 1.0$	$b_{20} + 5.5\,b_{21} + 1.0\,b_{22} + c_3 \leq 0.0$	$b_{30} + 4.5\,b_{31} + 1.0\,b_{32} + c_3 \leq 0.0$
$k=2$	4		$b_{20} + 2.5\,b_{21} + 3.0\,b_{22} - c_4 + 10\,d_4 \geq 1.0$	$b_{30} + 2.5\,b_{31} + 3.0\,b_{32} - c_4 \leq 0.0$
	5		$b_{20} + 2.0\,b_{21} + 1.5\,b_{22} - c_5 + 10\,d_5 \geq 1.0$	$b_{30} + 2.0\,b_{31} + 1.5\,b_{32} - c_5 \leq 0.0$
	6		$b_{20} + 6.5\,b_{21} + 4.5\,b_{22} - c_6 + 10\,d_6 \geq 1.0$	$b_{30} + 6.5\,b_{31} + 4.5\,b_{32} - c_6 \leq 0.0$
$k=3$	7		$b_{20} + 1.0\,b_{21} + 2.5\,b_{22} - c_7 \leq 0.0$	$b_{30} + 1.0\,b_{31} + 2.5\,b_{32} - c_7 + 10\,d_7 \geq 1.0$
	8		$b_{20} + 5.0\,b_{21} + 4.5\,b_{22} - c_8 \leq 0.0$	$b_{30} + 5.0\,b_{31} + 4.5\,b_{32} - c_8 + 10\,d_8 \geq 1.0$
	9		$b_{20} + 2.5\,b_{21} + 4.0\,b_{22} - c_9 \leq 0.0$	$b_{30} + 2.5\,b_{31} + 4.0\,b_{32} - c_9 + 10\,d_9 \geq 1.0$

Table 3. Classification Results for the Wine Recognition Data Set

From	Classified into (in-Sample)			Classified into (Validation)			Total
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	
Class 1	59	0	0	57	2	0	59
Class 2	0	71	0	1	68	2	71
Class 3	0	0	48	1	1	46	48

Table 4. Classification Results for the Iris Data Set

From	Classified into (in-Sample)			Classified into (Validation)			Total
	Iris setosa	Iris versicolor	Iris virginica	Iris setosa	Iris versicolor	Iris virginica	
Iris setosa	50	0	0	49	1	0	50
Iris versicolor	0	49	1	0	46	4	50
Iris virginica	0	0	50	0	2	48	50

Table 5. Classification Results for the MBA Admission Data Set

From	Classified into (in-Sample)			Classified into (Validation)			Total
	Admit	Not Admit	Borderline	Admit	Not Admit	Borderline	
Admit	30	0	1	30	0	1	31
Not admit	0	28	0	0	26	2	28
Borderline	0	0	26	1	1	24	26

Table 6. Data in Example 5

$k=1$				$k=1$				$k=1$			
i	$j=1$	$j=2$	$j=3$	i	$j=1$	$j=2$	$j=3$	i	$j=1$	$j=2$	$j=3$
1	6.26	17.02	11.25	11	5.64	16.10	10.80	21	5.86	18.39	13.87
2	5.81	18.57	14.19	12	8.44	16.37	15.07	22	6.37	18.43	12.27
3	9.81	19.30	12.42	13	6.33	17.55	14.03	23	6.14	15.51	13.11
4	8.06	17.69	11.26	14	7.39	19.43	11.74	24	6.32	18.07	16.07
5	8.15	15.60	11.63	15	7.81	19.97	15.34	25	8.30	18.60	14.32
6	6.42	17.45	14.91	16	6.89	15.64	14.77	26	10.19	18.95	14.37
7	6.78	17.86	13.63	17	8.33	16.50	12.60	27	5.36	18.66	11.20
8	6.40	17.39	14.63	18	6.45	18.22	12.16	28	7.59	15.25	11.16
9	6.19	15.57	13.69	19	8.47	18.80	11.10	29	8.87	15.87	12.07
10	7.38	17.44	11.99	20	5.51	15.31	11.99	30	6.26	16.16	11.16

Table 7. Classification Results for Example 5

From	Classified into (in-Sample)			Total
	Class 1	Class 2	Class 3	
Class 1	8	1	1	10
Class 2	3	5	2	10
Class 3	3	1	6	10