ИНСТИТУТ ПО МАТЕМАТИКА С ИЗЧИСЛИТЕЛЕН ЦЕНТЪР
INSTITUTE OF MATHEMATICS WITH COMPUTER CENTER

June 3, 1995

Abstract

A Combinatorial Approach to the Classification Problem

N. Yanev, S. Balev

БЪЛГАРСКА
АКАДЕМИЯ
НА НАУКИТЕ

BULGARIAN
ACADEMY
OF SCIENCES

Preprint

May 1995

No 11

Department of Operations Research

A combinatorial approach to the classification problem

N. Yanev, S. Balev

June 5, 1995

Abstract

We study the two-group classification problem which involves classifying an observation into one of two groups based on its attributes. The classification rule is a hyperplane which separates the groups optimally, i.e. the set of the points of group 1 (2) belonging to the halfspace allocated for the points of group 2 (1) is of minimal cardinality. Exact and heuristic algorithms for solving the problem are presented. Computational results are reported.

Introduction

The classification problem, also called a discriminant problem, involves classifying an observation into one of a number of mutually exclusive groups, according to its attributes. This is one of the most fundamental problems of scientific inquiry and has found application in fields as biology, artificial intelligence, social and administrative sciences.

Observations are characterized by their attributes, so they can be represented by vectors, each component of which is a value of some attribute. In the two-group linear discriminant analysis a hyperplane with a normal vector w and a bias w_0 is determined, so that an observation X is assigned to the first group if $Xw \leq w_0$ and to the second group otherwise. This hyperplane, also called a linear discriminant function, is determined from a training sample – a set of observations, which group membership is a priori known, so that it best separates the two groups according to some criterion. Different approaches to the classification problem are developed depending on this criterion.

In this paper we present an efficient algorithm for constructing a linear discriminant function which minimizes the number of misclassified observations from the training sample. This problem is known to be intractable, which necessitates developing efficient special purpose exact or heuristic algorithms.

We will use the following notations:

m – the number of attributes;

n – the number of observations (points) in the training sample;

$X_i = (x_{i1}, \dots, x_{im})$ – the i -th observation, $i \in \overline{1, n}$;

G_1 – index set for the points of the first group;

G_2 – index set for the points of the second group.

The rest of the paper is organized as follows. Section 1 briefly describes the most popular approaches to the classification problem, Section 2 presents more convenient equivalent form of the

problem. A branch-and-bound algorithm for solving the problem is described in section 3. Section 4 presents a set of heuristic algorithms. Section 5 summarizes the results of the conducted computational experiment.

1 Overview of approaches to the classification problem

Approaches to the classification problem can generally be separated into two groups – statistical and mathematical programming (MP) approaches.

Statistical approaches are the classical tool in discriminant analysis. In general, they operate from the assumption that each population has known distribution. Classification rules, produced by statistical approaches, minimize the probability of misclassification. When the population distributions are multivariate normal with mean vectors μ_i , a common variance-covariance matrix Σ , and relative frequencies π_i , $i = 1, 2$, the classification rule becomes:

Assign an observation X to G_1 if

$$X^T \Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) + \ln \frac{\pi_2}{\pi_1}$$

and to G_2 otherwise.

This is the most widely known Fisher's discriminant function [5]. When μ_1 , μ_2 , and Σ are replaced by their maximum likelihood estimators, a sample-based classification rule is obtained.

MP approaches are developed as an alternative to the statistical methods. As an example of linear programming (LP) approaches we will demonstrate the MMD (minimize maximum distance) model [6]:

Maximize d

subject to:

$$\begin{aligned} \sum_{j=1}^m x_{ij} w_j + d &\leq w_0 \text{ for } i \in G_1 \\ \sum_{j=1}^m x_{ij} w_j - d &\geq w_0 \text{ for } i \in G_2 \\ w_j, d &\text{ - unrestricted} \end{aligned}$$

This approach determines a hyperplane that maximizes the minimum distance, d , between an observation and the hyperplane.

After their introduction by Freed and Glover [6], [7], various LP formulations have been proposed [8], [9], [13] and shown to outperform statistical approaches, when the assumptions for group distributions are violated [4], [10], [12]. As well as their advantages, LP formulations have several defects [13]. One of them is that they may produce the trivial solution $w_j = 0$ for all j , i.e. meaningless classification rule. To overcome this defect, a normalization constraint $bw = c$ is added to the model, where b is a nonzero vector and c is a nonzero constant. The addition of this constraint prevents the zero solution, but causes some side effects as eliminating possible solutions, for which $bw = 0$. Sensitivity to the extreme values is also a problem of LP formulations. If an observation from one group is far within the region of the other group, LP models give

unsatisfactory results. This fault can be overcome by using mixed integer programming (MIP) formulations. For example:

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

subject to:

$$\begin{aligned} \sum_{j=1}^m x_{ij}w_j - Mz_i &\leq w_0 \text{ for } i \in G_1 \\ \sum_{j=1}^m x_{ij}w_j + Mz_i &\geq w_0 \text{ for } i \in G_2 \\ w_j &\text{ - unrestricted; } z_i \in \{0, 1\} \end{aligned}$$

where M is sufficiently large number. The binary variables z_i are 0 if the i -th observation is correctly classified and 1 otherwise. MIP formulations directly minimize the number of misclassified observations. Computational time for solving these formulations by general purpose methods, is significant. Bajgier and Hill [4] stated that “MIP approaches is unlikely to be practical for discriminant problems with more than 50 cases unless very efficient, special-purpose algorithms (either exact or heuristic) are developed”. Such algorithms are proposed in [1],[3],[11]. In the next sections we also present algorithms that minimize the number of misclassifications.

2 Problem formulation

Consider n points $X_i = (x_{i1}, \dots, x_{im})$, $i \in \overline{1, n}$ from two groups – G_1 and G_2 . The group membership of each point is known. The objective is to remove minimal number of them, so that the rest of them become linearly separable, i.e. the system

$$\begin{aligned} \sum_{j=1}^m x_{ij}w_j &\leq w_0 \text{ for } i \in G_1 \\ \sum_{j=1}^m x_{ij}w_j &\geq w_0 \text{ for } i \in G_2 \end{aligned}$$

has a nonzero solution after removing these points. It is proved in [2] that this problem, referred as MIN H-MISCLASSIFICATIONS, is NP -hard and moreover, very hard to approximate – not approximable within $2^{\log^{1-\epsilon} n}$ for any $\epsilon > 0$, unless $NP \subseteq DTIME(n^{\text{poly} \log n})$.

Let A be the matrix with columns

$$A_i = \begin{cases} (-1, X_i)^T & \text{for } i \in G_1 \\ (1, -X_i)^T & \text{for } i \in G_2 \end{cases}, i \in \overline{1, n}$$

and $w = (w_0, w_1, \dots, w_m)^T$. We prevent zero solution by normalization $bw > 0$ for an arbitrary nonzero vector b . So, we obtain the following problem:

(P) Find a set $I \subseteq \overline{1, n}$ of minimal cardinality, so that the system

$$\begin{aligned} A^T w &\leq 0 \\ bw &> 0 \end{aligned} \tag{1}$$

becomes feasible after removing constraints with numbers in I .

Consider the system

$$\begin{aligned} Ax &= b \\ x &\geq 0 \end{aligned} \tag{2}$$

According to the Farkas theorem, (1) is feasible if and only if (2) is infeasible. Hence, the initial problem **P** is equivalent to the following problem:

(R) Find a set $I \subseteq \overline{1, n}$ of minimal cardinality, so that the system (2_I) is infeasible.

where (2_I) is the system (2) in which the variables x_i for $i \in I$ are zeroed (removed). We will solve the problem **R** instead of the original problem **P**.

In this formulation, points on the discriminant hyperplane are considered as properly classified. Such a separation is called unstrict. When a strict separation is needed, we consider the system

$$\begin{aligned} \sum_{j=1}^m x_{ij} w_j &\leq w_0^1 \text{ for } i \in G_1 \\ \sum_{j=1}^m x_{ij} w_j &\geq w_0^2 \text{ for } i \in G_2 \end{aligned}$$

with a normalization constraint

$$w_0^2 - w_0^1 > 0$$

and construct a problem, similar to **R** in the same way.

From any optimal solution I^* of **R**, the separating hyperplane of the points, whose indices are out of I^* could be easily found. It is possible to solve some of the linear models with the rest of the points, already linearly separable, to obtain some secondary goal.

Further on we will use a basic representation of (2), namely

$$\sum_{i=1}^n y_i x_i = \bar{b} \tag{3}$$

where B is some non-singular submatrix of A of a full rank, $y_i = B^{-1}A_i$ and $\bar{b} = B^{-1}b$ are the values of the basic variables, i.e. the variables whose indices are in B . When $\bar{b} \geq 0$, we will call the relevant basis feasible. We will denote the j -th element of B by $B(j)$. The solution of (2) $x_{B(j)} = \bar{b}_j$ for $j \in \overline{0, m}$ and $x_i = 0$ for $i \notin B$ corresponds to the feasible basis B .

3 Exact algorithm

In this section we propose a branch-and-bound algorithm for solving the problem **R** to optimality.

3.1 Branching procedure

The branching strategy is based on the following theorem.

Theorem 1. Let I^* be an optimal solution of $\mathbf{R}$. Then:

- (i) for each $k \in I^*$, there exists a feasible basis $\mathcal{B}$ of (2), such that $k \in \mathcal{B}$;
- (ii) for each feasible basis $\mathcal{B}$ of (2), there exists $k \in I^*$, such that $k \in \mathcal{B}$.

PROOF.

(i) Let $k \in I^*$ and $I = I^* \setminus \{k\}$. The system (2_I) is feasible, because I^* is an optimal solution of $\mathbf{R}$. Let $\mathcal{B}$ be a feasible basis of (2_I) . Then (2_I) has the following basic representation

$$\sum_{i \in \mathcal{B}} y_i x_i + \sum_{i \in \overline{1, n} \setminus (I \cup \mathcal{B})} y_i x_i = \bar{b}.$$

Suppose that $k \notin \mathcal{B}$, then

$$\sum_{i \in \mathcal{B}} y_i x_i + \sum_{i \in \overline{1, n} \setminus (I \cup \mathcal{B} \cup \{k\})} y_i x_i = \bar{b}$$

is a basic representation of (2_{I^*}) , i.e. (2_{I^*}) is feasible, which contradicts the optimality of I^* .

(ii) Let $\mathcal{B}$ be a feasible basis of (2). Suppose that for each $k \in I^*$ $k \notin \mathcal{B}$. Then (2) has a basic representation

$$\sum_{i \in \mathcal{B}} y_i x_i + \sum_{i \in I^*} y_i x_i + \sum_{i \in \overline{1, n} \setminus (I^* \cup \mathcal{B})} y_i x_i = \bar{b}.$$

Hence, (2_{I^*}) has a basic representation

$$\sum_{i \in \mathcal{B}} y_i x_i + \sum_{i \in \overline{1, n} \setminus (I^* \cup \mathcal{B})} y_i x_i = \bar{b},$$

i.e. (2_{I^*}) is feasible. This contradiction proves (ii).

Let I be an index set. By $\mathbf{R}_I$ we note the problem $\mathbf{R}$, where the system (2) is replaced by (2_I) . We define a tree with problem $\mathbf{R}_I$ for some I in each node, in the following way:

1. $\mathbf{R}$ is the root of the tree.
2. Suppose $\mathbf{R}_I$ is in a certain node of the tree. Then:
 - (a) if the system (2_I) is infeasible, this node is a leaf of the tree;
 - (b) if the system (2_I) is feasible and $\mathcal{B}$ is its feasible basis, then the children of this node are the problems $\mathbf{R}_{I \cup \{B(j)\}}$ for $j \in \overline{0, m}$.

It follows from Theorem 1 that if I^* is an optimal solution of $\mathbf{R}$ and $|I^*| = s$, then $\mathbf{R}_{I^*}$ is a leaf on the s -th level of the tree.

So, the branching in any node will be performed on the basic variables of the corresponding problem.

It is useful to notice that one and the same problem may appear several times in the tree. For example, let k and l be in a basis of (2), k be in a basis of $(2_{\{l\}})$, and l be in a basis of $(2_{\{k\}})$. Then the problem $\mathbf{R}_{\{k, l\}}$ will appear two times at the second level of the tree. To avoid that, we proceed as follows. Let $\mathcal{B}$ be the feasible basis of problem $\mathbf{R}_I$, on which we perform branching. Then branching on $\mathcal{B}(0), \dots, \mathcal{B}(j-1)$ is not performed in the whole subtree with root $\mathbf{R}_{I \cup \{B(j)\}}$, even if they are basic in some node of this subtree. It is easy to see that in this way we avoid repeating problems without losing possible solutions.

When we pass from a basis of a certain node to a basis of its child, it is useful to leave as many of the old basic variables as possible in the new basis. This will increase the number of repeating problems and hence, will reduce the nodes of the tree.

3.2 Lower bounds

To obtain lower bounds of the solutions of the problems $\mathbf{R}_I$, we use Theorem 2, which immediately follows from Theorem 1.

Theorem 2. *Let $\mathcal{B}_1, \dots, \mathcal{B}_l$ be mutually exclusive feasible bases of the system (2) and I^* is an optimal solution of the problem $\mathbf{R}$. Then $|I^*| \geq l$.*

We find a sequence of mutually exclusive feasible bases of (2) by repeating the following steps:

1. Find a feasible basis of (2). If there is not such a basis, i.e. (2) is infeasible, stop.
2. Remove all basic variables from (2).

To find mutually exclusive bases of a certain problem, bases of its parent, already found, can be used.

We perform a depth-first branch-and-bound search using the described branching procedure. Lower bounds generated in any node are used to fathom in the usual way. In the next section we describe a heuristic used to determine an initial upper bound.

4 Upper bounds and heuristic algorithms

We propose methods that produce feasible solutions of the problem $\mathbf{R}$ close to the optimal solution. These methods are based on the following result.

Theorem 3. *Let (3) be an arbitrary basic representation of the system (2) and λ be a vector, such that $\langle \lambda, \bar{b} \rangle > 0$. Then the set $T = \{i \in \overline{1, n}, \langle \lambda, y_i \rangle > 0\}$ is a feasible solution of $\mathbf{R}$.*

PROOF. Multiplying (3) by λ we obtain

$$\sum_{i=1}^n \langle \lambda, y_i \rangle x_i = \langle \lambda, \bar{b} \rangle.$$

After removing the variables with numbers in T , the last equality becomes:

$$\sum_{\substack{i \in \overline{1, n} \\ \langle \lambda, y_i \rangle \leq 0}} \langle \lambda, y_i \rangle x_i = \langle \lambda, \bar{b} \rangle.$$

The left-hand side of the last equality is not positive for any values of the variables $x_i \geq 0$ while the right-hand side is positive by assumption. Hence, the system (2_T) is infeasible, i.e. T is a feasible solution of $\mathbf{R}$.

If we set all but j -th component of λ equal to zero, we obtain that the sets

$$T_j = \begin{cases} \{i \in \overline{1, n}, y_{ji} > 0\} & \text{if } \bar{b}_j > 0 \\ \{i \in \overline{1, n}, y_{ji} < 0\} & \text{if } \bar{b}_j < 0 \end{cases}$$

are feasible solutions of $\mathbf{R}$. Taking the minimal of these sets we obtain the simplest heuristic, U_1 :

U_1 is the set of minimal cardinality among the sets T_j for $j \in \overline{0, m}$ and $\bar{b}_j \neq 0$.

This idea can be generalized. Consider k rows from (3) with numbers $j \in J$, $|J| = k$:

$$\sum_{j=1}^n c_j x_j = c_0$$

where c_j are the vectors with components y_{ji} , $j \in J$, and c_0 is the vector with components $\bar{b}_j$, $j \in J$. We choose J so that $c_0 \neq 0$. Note that there are at least k linearly independent among the vectors c_i , $i \in I$ (for example, such are the unit vectors e_j , $j \in \overline{1, k}$). Let

$$T(\lambda) = \{i \in \overline{1, n}, \langle \lambda, c_i \rangle > 0\}.$$

By Theorem 3, if $\langle \lambda, c_0 \rangle > 0$ then $T(\lambda)$ is a feasible solution of **R**. Consider the following problem:

(**S_k**) Find a vector λ , such that $\langle \lambda, c_0 \rangle > 0$ and the set $T(\lambda)$ is minimal.

Obviously, any optimal solution of **S_k** is an upper bound for **R** and, as it will be shown later, the sharpness of the bounds increase with k . Next, We will describe an algorithm for solving that problem. Let

$$D(\lambda) = \{\mu : \langle \mu, c_0 \rangle > 0; \langle \lambda, c_i \rangle \langle \mu, c_i \rangle \geq 0, i \in \overline{1, n}\}.$$

Lemma 4. If λ is a solution of **S_k** then each vector from $D(\lambda)$ is also a solution of **S_k**.

PROOF. Let λ be a solution of **S_k** and $\mu \in D(\lambda)$. If $\langle \lambda, c_i \rangle \leq 0$ then $\langle \mu, c_i \rangle \leq 0$, i.e. if $i \notin T(\lambda)$ then $i \notin T(\mu)$. Then $T(\mu) \subseteq T(\lambda)$. But $\langle \mu, c_0 \rangle > 0$, hence μ is a solution of **S_k**.

For the sets $I \subseteq \overline{1, n}$, $|I| = k - 1$, for which the vectors c_i , $i \in I$ are linearly independent, we define rays d_I as follows:

$$\langle d_I, c_i \rangle = 0, i \in I \tag{4}$$

$$\langle d_I, c_0 \rangle > 0 \tag{5}$$

It is clear that for each set I there is at most one ray satisfying conditions (4) and (5).

Lemma 5. For each vector λ , such that $\langle \lambda, c_0 \rangle > 0$, there exists a set I such that $d_I \in D(\lambda)$.

PROOF. Consider the cone

$$D'(\lambda) = \{\mu : \langle \lambda, c_i \rangle \langle \mu, c_i \rangle \geq 0, i \in \overline{1, n}\}.$$

Let $d_1, \dots, d_s$ be all of the extreme rays of $D'(\lambda)$. $\lambda \in D'(\lambda)$, then

$$\lambda = \sum_{i=1}^s \alpha_i d_i,$$

where $\alpha_i \geq 0$, $i \in \overline{1, s}$. Hence,

$$\langle \lambda, c_0 \rangle = \sum_{i=1}^s \alpha_i \langle c_0, d_i \rangle > 0.$$

Then $\langle c_0, d_i \rangle > 0$ for at least one $i \in \overline{1, s}$. Let d be one of the rays $d_1, \dots, d_s$, such that $\langle c_0, d \rangle > 0$. Then $d \in D(\lambda)$. d is a section of $k - 1$ hyperplanes

$$\{\mu : \langle \lambda, c_i \rangle \langle \mu, c_i \rangle = 0\}.$$

Let I be the set of the numbers of these hyperplanes. Since d satisfies (4) and (5), $d \equiv d_I$.

As an immediate corollary from Lemma 4 and Lemma 5 appears the following result.

Theorem 6. *Some of the rays d_I , defined by (4) and (5) is an optimal solution of the problem S_k .*

The rays d_I are easily found by (4) and (5). By checking all of them we can find a solution of S_k .

Thus we obtained a method for finding upper bounds U_k , $k \in \overline{1, m+1}$ by considering k rows of (3) and solving the problems S_k . It is easy to see that $|U_1| \geq \dots \geq |U_{m+1}|$, but the computational effort for finding U_k is considerable for big k . These heuristics are useful only for small k . We will prove that as is to be expected, U_{m+1} produces the optimal solution of R . For that purpose, note that the problem S_{m+1} can be defined as follows:

Find the minimal among the sets $T(\lambda) = \{i \in \overline{1, n}, \langle \lambda, y_i \rangle > 0\}$ where $\langle \lambda, \bar{b} \rangle > 0$.

and the initial problem P can be written down in the following way:

Find the minimal among the sets $I(w) = \{i \in \overline{1, n}, \langle w, A_i \rangle > 0\}$ where $\langle w, b \rangle > 0$.

Lemma 7. *The problems S_{m+1} and P are equivalent (have the same solutions).*

PROOF. Let $T(\lambda^0)$ be an optimal solution of S_{m+1} and $w^0 = (B^{-1})^T \lambda^0$. Then:

$$\langle w^0, b \rangle = \langle (B^{-1})^T \lambda^0, b \rangle = \langle \lambda^0, B^{-1} b \rangle = \langle \lambda^0, \bar{b} \rangle > 0,$$

$$\langle w^0, A_i \rangle = \langle (B^{-1})^T \lambda^0, A_i \rangle = \langle \lambda^0, B^{-1} A_i \rangle = \langle \lambda^0, y_i \rangle, \quad i \in \overline{1, n}.$$

Hence, $I(w^0)$ is a feasible solution of P and $I(w^0) \equiv T(\lambda^0)$. Suppose that $I(w^0)$ is not an optimal solution of P . Then there exists w^1 , such that $\langle w^1, b \rangle > 0$ and $|I(w^1)| < |I(w^0)|$. Let $\lambda^1 = B^T w^1$. Then:

$$\langle \lambda^1, \bar{b} \rangle = \langle B^T w^1, B^{-1} b \rangle = \langle w^1, B B^{-1} b \rangle = \langle w^1, b \rangle > 0,$$

$$\langle \lambda^1, y_i \rangle = \langle B^T w^1, B^{-1} A_i \rangle = \langle w^1, B B^{-1} A_i \rangle = \langle w^1, A_i \rangle > 0, \quad i \in \overline{1, n}.$$

Hence, $T(\lambda^1)$ is a feasible solution of S_{m+1} and $T(\lambda^1) \equiv I(w^1)$. But $|T(\lambda^1)| = |I(w^1)| < |I(w^0)| = |T(\lambda^0)|$, which contradicts the optimality of $T(\lambda^0)$. Therefore, $I(w^0) \equiv T(\lambda^0)$ is an optimal solution of P .

It can be proved that an optimal solution of P is also an optimal solution of S_{m+1} in the same way.

From the last lemma and the equivalence between the problems P and R , it follows that:

Theorem 8. *U_{m+1} is an optimal solution of the problem R .*

The heuristics U_k depend on the basis where they are computed, i.e. $U_k = U_k(\mathcal{B})$. The next theorem shows the strong influence of the basis upon these heuristics.

Theorem 9. *There exists a basis $\mathcal{B}$ of the system (2), such that $U_1(\mathcal{B})$ is an optimal solution of the problem R .*

PROOF. Let (3) be an arbitrary basic representation of (2) and λ be a solution of the problem S_{m+1} . Due to Theorem 8, $T(\lambda) \equiv U_{m+1}$ is an optimal solution of R . According to Theorem 6, there exists a set I , such that $|I| = m$; the vectors y_i , $i \in I$, are linearly independent; $\langle \lambda, y_i \rangle = 0$, $i \in I$; and $\langle \lambda, \bar{b} \rangle > 0$. We construct a matrix M with the following columns: $M_1, \dots, M_m$ are

the vectors $y_i, i \in I$, and $M_0 = y_s$, where we choose s so that the matrix M is non-singular. By multiplying (3) from left by M^{-1} , we obtain a new basic representation

$$\sum_{i=1}^n y'_i x_i = \bar{b}'$$

where $y'_i = M^{-1}y_i$, $\bar{b}' = M^{-1}\bar{b}$. The basis corresponding to this representation is $\mathcal{B} = \{s\} \cup I$. We will prove that $\mathcal{B}$ is the required basis. Let $\lambda' = M^T\lambda$. Then:

$$\langle \lambda', \bar{b}' \rangle = \langle M^T\lambda, M^{-1}\bar{b} \rangle = \langle \lambda, MM^{-1}\bar{b} \rangle = \langle \lambda, \bar{b} \rangle > 0,$$

$$\langle \lambda', y'_i \rangle = \langle M^T\lambda, M^{-1}y_i \rangle = \langle \lambda, MM^{-1}y_i \rangle = \langle \lambda, y_i \rangle, i \in \overline{1, n}.$$

Hence, $T(\lambda') \equiv T(\lambda)$ is the optimal solution of **R**. $y'_i, i \in I$ are unit vectors and moreover, $\langle \lambda', y'_i \rangle = \langle \lambda, y_i \rangle = 0, i \in I$. Hence, $\lambda'_j = 0, j \in \overline{1, m}$. Since $\langle \lambda', \bar{b}' \rangle > 0$, $\lambda'_0 \neq 0$ and has the same sign as $\bar{b}'_0$. Then

$$T(\lambda') = \{i \in \overline{1, n}, \langle \lambda', y'_i \rangle > 0\} = \{i \in \overline{1, n}, \lambda'_0 y'_{0i} > 0\} = T_0$$

is an optimal solution of **R**. U_1 was defined to be the minimal among the sets $T_0, \dots, T_m$, hence $U_1 \equiv T_0$. The last means that the heuristic U_1 , applied for the new basis $\mathcal{B}$, produces the optimal solution of the problem **R**.

Using the last result, a local-search algorithm suggests itself. Let $f(\mathcal{B}) = |U_1(\mathcal{B})|$ and a neighbourhood of $\mathcal{B}$ be the set $N(\mathcal{B})$ of all bases that differ from $\mathcal{B}$ only by one element (neighbours of $\mathcal{B}$ in the usual sense). We minimize locally the function f using the neighbourhood system N . As starting points we use the mutually exclusive bases of (2), found to derive a lower bound of **R**. This heuristic is used to determine an initial lower bound for the branch-and-bound algorithm.

5 Computational experiment

In this section we describe the computational testing done to evaluate the performance of the proposed branch-and-bound (BB) algorithm and the local search (LS) heuristic. The discriminant power of the MIP approaches to the classification problem was investigated in detail in [4], [11], [14]. That is why we focused our study only on computational efficiency of our algorithms. For comparative purpose, we used a part of Joachimsthaler and Stam [10] simulation data experimental design also used in [11] and [1]. We generated three-variate normal distributions with different means and dispersion matrices. Different settings of these parameters form three cells. Table 1 shows the population parameters for each cell. I is the unit matrix and $e = (1, 1, 1)^T$.

Table 1: Parameter settings.

Cell	Dispersion matrices		Mean vectors	
	G_1	G_2	G_1	G_2
1	I	I	0	.5e
2	I	$2I$	0	.6e
3	I	$4I$	0	.8e

For each cell, 100 random samples were generated. Each sample contained 100 observations – 50 observations from G_1 and 50 observations from G_2 . We compare our results with the results of Koehler and Erenguc [11] for their exact algorithm PMM and heuristic BPMM. Note however that our samples are not exactly the same as used in [11], they only have the same parameters. Table 2 shows the average misclassification rates for each cell of the study. It is seen in this table the surprisingly good performance of LS.

Table 2: Average misclassification rates.

Cell	BB	LS	PMM	BPMM
1	.2640	.2648	.258	.266
2	.2529	.2533	.255	.262
3	.2194	.2197	.220	.226

The closeness of LS to the optimal solution becomes more obvious from table 3. It shows the number of the problems in which LS finds extra misclassifications (0,1 or 2) with respect to the optimal one found by the exact BB algorithm. Only in 14 from 300 generated samples LS differs from the optimum. In no case LS misclassified more than two observations above the optimal one.

Table 3: Number of misclassifications of LS over BB.

Cell	0	1	2
1	93	6	1
2	96	4	0
3	97	3	0

Tables 4 through 8 summarize the statistics for computational effort required. Table 4 shows the number of the nodes of the tree generated by BB. Table 5 shows the number of linear programs (LPs) solved by BB (by LP we mean finding of a feasible basis of a system (2_I) for some I). Note that in a certain node of the tree, more than one LP may be solved because of the searching of lower bounds. Tables 6 and 7 show the number of pivots required by BB and LS respectively. The number of pivots required by BB includes the number of pivots required by LS because LS was used to obtain an initial upper bound for BB.

Table 4: Number of nodes of the tree generated by BB.

Cell	Average	St. deviation	Minimum	Maximum
1	8,532	10,774	174	62,298
2	7,072	9,134	81	44,538
3	2,405	2,880	47	19,816

Table 5: Number of LPs solved by BB.

Cell	Average	St. deviation	Minimum	Maximum
1	19,734	25,280	417	143,972
2	16,041	21,052	242	99,744
3	5,327	6,565	115	45,679

Table 6: Number of pivots performed by BB.

Cell	Average	St. deviation	Minimum	Maximum
1	137,377	159,238	13,581	897,485
2	101,778	109,990	12,453	507,973
3	39,849	33,016	10,313	236,037

Table 7: Number of pivots performed by LS.

Cell	Average	St. deviation	Minimum	Maximum
1	13,538	1,703	8,712	17,400
2	12,834	1,712	8,705	17,667
3	11,283	1,383	8,353	14,287

As it is seen in these tables, the computational effort required by BB sharply decreases when the group overlap is smaller. For example, when the average misclassification rate decreases only by 0.335 (cells 2–3), the average number of pivots is approximately 2.5 times lower. The computational effort required by LS is relatively constant and significantly lower.

For comparison, table 8 shows some average results from the study of Koehler and Erenguc for their exact algorithm. Linear programs solved by their algorithm have a matrix of the same size as ours but they also have an objective function and a simple lower bound for each variable. A minor pivot is one in which a nonbasic variable switches its boundary value. A major pivot is one where a nonbasic variable is pivoted into basis. Major pivots are computationally more expensive than the minor ones. As there are no simple lower bounds for the variables in our models, we perform only major pivots.

Table 8: Some average results from the study of Koehler and Erenguc.

Cell	LPs	Minor pivots	Major pivots
1	99,417	229,612	62,720
2	100,284	233,216	64,437
3	61,821	161,663	53,485

The results of the conducted computational experiment characterize the proposed branch-and-bound algorithm as very efficient algorithm for minimizing the number of the misclassifications in the training sample especially when the group overlap is relatively small. The local search heuristic produces solutions, very close to the optimum and requires significantly lower computational effort, so it is an valuable alternative to the MIP approaches.

Acknoledgments

The authors wish to thank dr. Ognyan Asparouchov for stimulating and helping us to start study this problem. Special thanks are given to dr. Vladimir Tzokov for the help in solving the problems S_2 .

References

- [1] ABAD, P. and BANKS, W. New LP based heuristics for the classification problem. *European Journal of Operational research*, 1993, **67**, 88-100.
- [2] AMALDI, E. and KANN, V. On the approximability of removing the smallest number of relations from linear systems to achieve feasibility. 1994, Unpublished manuscript.
- [3] BANKS, W. and ABAD, P. An efficient optimal solution algorithm for the classification problem. *Decision sciences*, 1991, **22**, 1008-1023.
- [4] BAJGIER, S. and HILL, A. An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences*, 1982, **13**, 604-618.
- [5] FISHER, R. The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, **7**, 179-188.
- [6] FREED, N., and GLOVER, F. A linear programming approach to the discriminant problem. *Decision Sciences*, 1981, **12**, 68-74.
- [7] FREED, N. and GLOVER, F. Simple but powerful goal programming models for discriminant problems. *European Journal of Operational research*, 1981, **7**, 44-60.
- [8] FREED, N. and GLOVER, F. Evaluating alternative linear programming models to solve the two-group discriminant problem. *Decision sciences*, 1986, **17**, 151-162.
- [9] GLOVER, F., KEENE, S., and DUEA, B. A new class of models for the discriminant problem. *Decision sciences*, 1988, **19**, 269-280.
- [10] JOACHIMSTHALER, E. and STAM, A. Four approaches to the classification problem in discriminant analysis: an experimental study. *Decision sciences*, 1988, **19**, 322-333.
- [11] KOEHLER, G. and ERENGUC, S. Minimizing misclassifications in linear discriminant analysis. *Decision Sciences*, 1990, **21**, 63-85.
- [12] MARKOWSKI, C. and MARKOWSKI, E. An experimental comparison of several approaches to the discriminant problem with both qualitative and quantitative variables. *European Journal of Operational research*, 1987, **28**, 74-78.
- [13] RAGSDALE, C. and STAM, A. Mathematical programming formulations for the discriminant problem: An old dog does new tricks. *Decision sciences*, 1991, **22**, 296-307.
- [14] STAM, A. and JOACHIMSTHALER, E. A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational research*, 1990, **46**, 113-122.