BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **6**, No 2

Sofia • 2006

The Two-Group Classification Problem Challenge – Polynomial Heuristic Algorithms

Vassil G. Guliashki

Institute of Information Technologies, 1113 Sofia

Abstract: *The paper presents a heuristic approach for development of polynomial-time algorithms, solving the two-group classification problem. The proposed algorithm ALS, based on this heuristic approach, has been tested on 200 test problems with 6 attributes and 150 training sample observations (75 per group), and with 10% to 30% overlapping of both groups in the training sample. The obtained results are compared with that one, obtained by means of three other heuristic algorithms, one exact algorithm and one statistical method on the same test problems. The computational complexity of ALS algorithm is very encouraging.*

Keywords: *Discriminant analysis, two-group classification, polynomial heuristic algorithms.*

1. Introduction

The two-group classification problem is very important, because it appears in many business areas (economics, marketing, finance and management), as well as in engineering, medical and social sciences.

Let two groups of objects g_1 and g_2 be given. There are available $m = m_1+m_2$ observations of these objects in a training sample (m_1 are from g_1, and m_2 are from g_2). The objects are described by an n-component vector of attributes $x = (x_1, ..., x_n)$. Usually $m_1 > n$ and $m_2 > n$. The objective of discriminant analysis is to find a function (classifier) $f(x,w)$ separating the two groups (classifying the objects to the corresponding groups).

Different type approaches have been used over the years to solve this problem and corresponding methods are developed. Many investigations have been devoted to the statistical (nonparametric and parametric) methods. The most widely known statistical methods, such as Fisher's linear discriminant function (LDF) [7], Smith's quadratic discriminant function (QDF) [23] and the logistic discriminant function

(LGD) [5] may yield poor classification results if the data sets of the both groups are highly skewed or scattered and the training sample observations are contaminated by outliers (see [3, 20]). The LDF-methods have robust classification accuracy for normally distributed data sets and perform poorly if the deviations from normality are significant. The logistic regression method is a parametric statistical method. Also the LDF- and QDF-methods are parametric statistical methods, using a distance measure, based on L_2-norm. However it is well known that criteria based on higher norm distances perform poorly if extreme training sample observations are available (see [6]). For this reason many researchers have directed their efforts to develop nonparametric classification methods. There exist nonparametric Mathematical Programming (MP) methods minimizing the absolute distances to the hyperplane that separates the groups (L_1-norm-based methods) and MP methods, which minimize the actual number of misclassified observations (L_0-norm-based methods). In [21] is supported the use of MP methods, because they do not make any assumptions about the distributional characteristics of the attribute populations. These methods focus on the search space region, where overlap of the groups occurs. An L_1-norm-based method in [9] proposes the MSD (minimize the sum of deviations) model. Some experimental results [15] show that the MSD method performs more accurately than the LDF and QDF methods. An L-norm-based method [8] minimizes the maximum deviation (MMD). Other known L_1-norm-based method, solving the two-group classification problem, is the method optimizing the sum of deviations (OSD) [3, 19], as well as the Hybrid method [11]. The L_1- and L_∞-norm methods can be realized by means of linear programming (LP) techniques.

Methods, which use another MP approach, are the mixed-integer programming (MIP) methods. They minimize the number of misclassified training sample observations directly. The MIP methods can be viewed as L_p-norm methods with $p \to 0$ and are referred as MP-L_0 methods (see [2]). Methods belonging to this group are those, proposed in [2, 17, 25]. Unfortunately the MIP problems are proven to be NP-hard (see [10, 15]). The exact methods to solve these problems have exponential computational complexity. In the concrete case the computational efforts in the exact methods increase exponentially as a function of the training sample size and of the number of attributes. For this reason some exact algorithms which take advantage of the special structure and characteristics of the problem formulation are developed (see [4, 2, 25]) and the Divide and Conquer (D&C) algorithm [6]). To solve large-size two-group classification problems with a relative small deviation from optimality some researchers have developed heuristic algorithms, which drastically reduce the computational efforts in comparison to the exact algorithms (see [1, 12, 13, 14, 17, 22]). Relationships between support vector machines and the generalized linear discriminant analysis applied to the support vectors are studied in [16]. In this connection, exact generators of random vectors are proposed in [18, 19]. Illustrating the relationship, it is shown in [16] that the classification problem can be interpreted as a data reduction problem.

In this paper a heuristic approach and the possible development of polynomial algorithms based on it are considered. The paper is organized as follows: In Section 2 a brief formulation of the problem is presented. Section 3 states the mentioned heuristic approach and a possible basic algorithm. In Section 4 the performance of different exact and heuristic algorithms is compared. In Section 5 an illustrative example is presented. Some conclusions are made in Section 6.

2. The problem formulation

Let we consider two groups of objects g_1 and g_2 and $x_i = (x_1, ..., x_n)$, $i = 1, ..., m$, are training sample observations, such that m_1 of them belong to g_1 and m_2 of them belong g_2. Hence $m = m_1+m_2$ and x_i, $i = 1,..., m$, are n-dimensional vectors of attributes.

The most frequently used classifier is linear. The objective in this paper is to find an n-dimensional hyperplane in the attribute space $f(x, w) = x^Tw$, such that $x \in g_1$, if $f(x, w) \geq w_0$, otherwise $x \in g_2$. Here w is n-dimensional vector, containing the parameters of the classifier (the coefficients of the separating hyperplane), and w_0 is the cut-off value. The problem is formulated as a Mixed-Integer Programming problem (MIP-problem) as follows:

$$\min \ z = \sum_{i=1}^{m} \delta_i \tag{1}$$

subject to

$$-x_i^Tw + w_0 + M\delta_i \geq 0, \ i \in g_1, \ i = 1,..., m_1, \tag{2}$$

$$x_i^Tw - w_0 + M\delta_i \geq \varepsilon, \ i \in g_2, \ i = 1,..., m_2, \tag{3}$$

where w_k, $k = 0, 1,..., n$, are unrestricted in sign real variables, each of the binary variables δ_i correspond to one observation in the training sample, so that $\delta_i = 1$, if the i-th observation is misclassified and $\delta_i = 0$ if the i-th observation is correctly classified. The objective function value z is equal to the number of misclassifications. M is a sufficiently large, and ε is sufficiently small positive real number, for example $\varepsilon = \sqrt[3]{\text{macheps}}$, where "macheps" is the computer's machine precision.

3. Heuristic approach

Elements of the heuristic approach, discussed here has been proposed in [12, 13]. These ideas are developed further here and are illustrated by a test example in Section 5, in order to become clearer.

From geometrical point of view one training sample observation is one point in the n-dimensional Euclidean space of the attributes. In case the matrix of n arbitrary chosen different x_i^T, where $i = 1, ..., n$; training sample vectors (points) is nonsingular, i. e. it is of full rank, the chosen n points determine an unique hyperplane in this n-dimensional space. Let we assume that every one such matrix is nonsingular, i. e. the Haar condition holds (see [25]). Then each combination of n training sample observations (points) will determine an unique hyperplane. This assumption doesn't decrease the generality of the consideration, because the choice of a training sample corresponding to this condition is not difficult. The problem (1)-(3) is a combinatorial one. There are C_m^n hyperplanes, defined by all possible combinations of the training sample observations. The optimal hyperplane, separating the groups g_1 and g_2 with a minimum number of misclassifications may be obtained by the complete enumeration of all these hyperplanes. Usually the problem (1)-(3) has not an unique solution, i. e. there are several optimal hyperplanes. By means of some heuristic techniques the number of enumerated hyperplanes may be drastically reduced.

To obtain the coefficients of the hyperplane determined by the n training sample observations x_i^T, a determinant is solved:

$$(4)\qquad \begin{vmatrix} (u_1 - x_{11}), & (u_2 - x_{12}), & \ldots, & (u_n - x_{1n}) \\ (x_{21} - x_{11}), & (x_{22} - x_{12}), & \ldots, & (x_{2n} - x_{1n}) \\ \ldots & \ldots & \ldots & \ldots \\ (x_{n1} - x_{11}), & (x_{n2} - x_{12}), & \ldots, & (x_{nn} - x_{1n}) \end{vmatrix} = w_1 u_1 + w_2 u_2 + \ldots + w_n u_n - w_0 = 0,$$

where $(x_{i1}, x_{i2}, \ldots, x_{in}) = x_i^{\mathrm{T}}$, $u_1, u_2, \ldots, u_n$ are variables in the attribute space and w_k, $k = 0, 1, \ldots, n$) are the coefficients of the concrete hyperplane. The obtained coefficients w_k may be substituted in the system (2)-(3) and each violated inequality i in this system will lead to $\delta_i = 1$, so that the objective function z in (1) will increase by 1.

Grounded on geometrical reasons an initial hyperplane may be constructed, so that it separates relatively well the two groups of objects (observations). For example the weight centers O_1 and O_2 (mean vectors) correspondingly for the groups g_1 and g_2 may be calculated:

$$(5)\qquad O_{1j} = \left(\sum_{i=1}^{m_1} x_{ij} \right) / m_1, \quad i \in g_1, \; i = 1, \ldots, m_1, \; j = 1, \ldots, n;$$

$$(6)\qquad O_{2j} = \left(\sum_{i=1}^{m_2} x_{ij} \right) / m_2, \quad i \in g_2, \; i = 1, \ldots, m_2, \; j = 1, \ldots, n;$$

Then the normal vector h of the initial hyperplane H_0 may be chosen as

$$(7)\qquad h = O_2 - O_1, \text{ or } h_j = O_{2j} - O_{1j}, \quad j = 1, \ldots, n;$$

The initial hyperplane H_0 may pass through a chosen point (observation) x^{T} of the training sample or through the point $O = (O_2 - O_1)/2$. If a point x^{T} is chosen, then the coefficients of the hyperplane H_0 are determined as follows:

$$(8)\qquad w_j = h_j; \; j = 1, \ldots, n; \; w_0 = - x^{\mathrm{T}} h.$$

A simple way to construct a good initial hyperplane H_0 is to construct a hyperplane with normal vector h through each point x_i^{T}, $i = 1, \ldots, m$; in the training sample, to calculate the z-value for each of these hyperplanes and then to choose as initial hyperplane that one, corresponding to the minimum z-value. During this experiment it may occurs for some hyperplanes, that more than the half of the inequalities (2)-(3) are violated. In this case the hyperplane coefficients should be taken with the opposite sign: $w_j = - w_j$, $j = 0, \ldots, n$.

After the initial separating hyperplane H_0 is found, it is not difficult to find the n closest situated to it (in Euclidean sense) training observations (points). They define a new hyperplane, which may be denoted by H_1.

Having a hyperplane H defined by n points x_i^{T} from the training sample a set S_H of their indices i may be constructed. The replace of one corresponding to $i \in S_H$ point by another point from the training sample, which index doesn't belong to S_H will lead to a new hyperplane, slightly turned in comparison to the former one. The new hyperplane may (or may not) separate better the groups g_1 and g_2. Intuitively there are points in the training sample, having a great probability to take part in the definition of the optimal hyperplane H^*. Let the search of such points is focused in the region, where the both groups g_1 and g_2 overlap. A set P of indices of the training sample points, which probably will take part in the definition of the optimal hyperplane H^* may be created. It will include the closest situated k points ($k/2$ from g_1 and $k/2$

from g_2) to the current hyperplane H. Here k is even number. These points are considered as "promising" observations. Another criterion may be the value $d_i = |x_i^T w|$. The observations having value d_i closest to 0 may be considered as "promising". Let the training sample points are arranged according to their Euclidean distance to the current hyperplane H or according to their d_i -values. The first k of them are included in the set P.

The main idea of the heuristic approach presented here is to perform iterative enumeration of hyperplanes in such manner, that at each iteration a set of hyperplanes is constructed, replacing each point with index in S_H by the point x_j^T, where $j \in P$. This means that at each iteration n new hyperplanes will be enumerated. One enumeration cycle for the set S_H of the current hyperplane H will perform k iterations, because there are k indices in P: $j = 1, ..., k$. Hence the attempt to obtain a better separating hyperplane on the basis on the current hyperplane H will cost by such enumeration kn calculations of the determinant (4) and kn check-ups of the system (2)-(3) to calculate the z-value. In case n is a great number, for example if $n>20$, it is recommended to be taken $k = 10$, i. e. only 5 "promising" points from each group g_1 and g_2 will be included in P. The possible heuristic may perform several iteration cycles improving the z-value until there is no more improvement of z-value during the last iteration cycle.

Another idea is the information of the history of the search process to be used. In case the observation indices, defining *kbest* last found hyperplanes, arranged according their corresponding z-values, starting with the minimal one, are stored in an array MINH, a rating of each training sample observation may be calculated as follows: Let the observation x_j^T takes part in q_j hyperplanes from the hyperplanes, stored in MINH. Let Q_j is the index-set of these hyperplanes. Then the observation x_j^T has a rating: $\text{rate}(x_j^T) = \left(\sum_{i \in Qj} z_i \right) / q_j$. In case $q_j = 0$, $\text{rate}(x_j^T) = \min(m_1, m_2)$. Then the observations indices may be arranged in a list according the ratings of the observations, starting with the minimum $\text{rate}(x_j^T)$. A new set of "promising" observations P' may be created, containing k' indices among the first in the obtained rating list, which have not been included in the set S_H. An enumeration cycle may be performed by the indices in P′ like the enumeration cycle by P.

After all performed enumerations the two closest points (observations) to the current best found hyperplane, but not lying on it, may be selected (the one from group g_1 and the other from group g_2). The hyperplanes defined by all combinations of these two points with the points, which indices are included in S_H may be enumerated. If the number of attributes $n = 10$, then $C_n^2 = C_{10}^2 = n(n-1)/2 = 45$ combinations (new hyperplanes) will be enumerated.

To complete the search process a new enumeration cycle may be performed using the best found hyperplane and the corresponding set P.

At the end a simple transformation of the optimal hyperplane should be performed. This is necessary, because here it was implicitly assumed, that all training sample observations lying on the separating hyperplane are correctly classified. Usually this is not true, taking into account that the set S_H could contain indices of observations from both groups g_1 and g_2. In this case the best found hyperplane should be turned slightly, so that each of the training observations on it go to the corresponding group and the other $m - n$ observations keep their positions in the corresponding subspace. The transformation of the best-found hyperplane should be performed as follows:

Until this moment the found vector w satisfies the system

(9) $$-x_i^{\mathrm{T}} w + w_0 = 0, \quad i \in S_H \text{ and } i \in g_1,$$

or

(10) $$x_i^{\mathrm{T}} w - w_0 = 0, \quad i \in S_H \text{ and } i \in g_2.$$

Let $\hat{w}$ satisfies the system (2)-(3). It should be found the correction $\bar{a}$ in $\hat{w} = w + \bar{a}$, where $\hat{w}$ and $\bar{a}$ are n-dimensional vectors. Taking into account that ε is a small positive constant (see Section 2), and the system (9)-(10), then $\bar{a}$ should satisfy:

(11) $$-x_i^{\mathrm{T}} \bar{a} \geq 0, \quad i \in S_H \text{ and } i \in g_1,$$

(12) $$x_i^{\mathrm{T}} \bar{a} \geq \varepsilon, \quad i \in S_H \text{ and } i \in g_2.$$

Considering (11)-(12) as equalities $\bar{a}$ can be calculated by means of the following formula

(13) $$\bar{a} = X^{-1}[\varepsilon],$$

where X is a $n \times n$ matrix, which rows are $-x_i^{\mathrm{T}}$, $i \in S_H$ and $i \in g_1$; and x_i^{T}, $i \in S_H$ and $i \in g_2$; $[\varepsilon]$ is an n-dimensional vector, which i-th component is equal to 0 if $i \in S_H$ and $i \in g_1$; otherwise (if $i \in S_H$ and $i \in g_2$) it is equal to ε.

In case the training sample observations, which indices are in S_H, belong only to group g_2, then simply w_0 is changed: $\hat{w}_0 = w_0 - \varepsilon$.

In case the observations, which indices are in S_H, belong only to group g_1, the coefficients of the obtained hyperplane don't need any change.

4. Algorithmic scheme ALS

Step 1. Calculate the coefficients of an initial hyperplane H_0, solving (5), (6), (7) and (8). Choose among the given m observations the best observation, through which passes H_0.

Step 2. Calculate the coefficients of the hyperplane H_1, passing trough the n closest to H_0 training sample observations.

Step 3. Create set S_H, containing the indices of observations defining H_1. Create the set P, containing the first k indices of the closest observations (points) to H_1.

Step 4. Perform an enumeration cycle based on indices in P:

Let $j=0$ and $i=0$.

$i = i+1$ ***While*** $i \leq k$ **do:**

$j = j+1$ ***While*** $j \leq n$ **do:**

Replace x_j^{T}, $j \in S_H$ by x_i^{T}, $i \in P$.

Calculate the coefficients of the hyperplane

defined by the indices in S_H, solving **(4)**.

Endwhile

Endwhile

Let by H_{best} is denoted the best obtained hyperplane by this enumeration cycle. Update the set S_H by the indices of H_{best}.

Step 5. Calculate the rating of each training sample observation and rearrange the indices of the observations, starting with that one, having minimum rating value. Create a new set P', containing the first k' indices in the so obtained list of indices.

Step 6. Perform an enumeration cycle like that one in Step 4, but based on indices in P'.

Step 7. Find the two closest observations (points) to the current best found hyperplane, but not lying on it (the one from group g_1 and the other from group g_2). Create the set P'' with the indices of these two points.

Step 8. Perform an enumeration cycle like that one in Step 4, but based on indices in P''.

Step 9. Perform an enumeration cycle like that one in Step 4**,** based on indices in *P.* Repeat this step until no more new hyperplanes with *z*-values equal or less to the *z*-value of the best obtained hyperplane (before this step) are generated.

Step 10. If it is necessary, change the cut-off value of the final hyperplane H_{best} or turn slightly the hyperplane H_{best} to classify correctly the observations lying on it and obtain the hyperplane H_{best}', calculating its coefficients as in (13).

END of the calculations.

The algorithmic scheme, presented above is open for further improvement by other new ideas for creating of set *P*, and for including of such number of indices in it, that is the most appropriate for the size and properties of the solved problem.

Step 1 with the check of m training sample observations to find out through which of them passes H_0, Step 7 and Step 8 with the full enumeration of C_n^2 hyperplanes are new. The algorithm ALS may be extended by other new steps, based on the same heuristic approach. Similar algorithms are proposed in [12, 13] and test results are presented in [12, 13, 14].

5. Comparison between different heuristic algorithms and one exact algorithm

To evaluate the relative deviation from optimality of the obtained solutions in [12, 13, 14] the following formula is used:

$$\text{err} = \left(\left(\sum_{i=1}^{N} z_i - z_i^* \right) .100\% \right) / N.m, \tag{14}$$

where N is the number of the test problems, z_i and z_i^* are the best found and the optimal *z*-value for the *i*-th test problem and $m = m_1 + m_2$ is the number of training sample observations.

Eight data sets, each one containing 25 test problems are used to test the presented algorithm. The problems are with 6 attributes and 150 observations in two groups (75 per group). The problems in the different data sets are randomly generated by means of different mean vectors and covariance matrices, so that the both groups overlap to a different degree. The overlapping varies among the data sets between 10% and 30%. By means of (14) the obtained mean error for these 200 test problems is evaluated: err = 4.5%. Similar results are obtained in [14] on the same test problems. In Table 1 are presented the test results for these 200 test problems about the mean error and the mean arithmetical operations for one test problem of 4 heuristic algorithms - HG1, HG2 and FCS from [14] and ALS, presented here, one exact – D&C from [6] (known as one of the best exact algorithms for this class of problems) and one statistical method – LDF method.

Procedure	**HEURISTIC ALGORITHM / METHOD**					
	HG1	HG2	FCS	ALS	LDF	D&C
Mean error [%]	11.53	6.27	4.55	4.50	7.98	0.00
Math. operations	$\approx 48.n^3$	$\approx 350.n^3$	$\approx 808.n^3$	$\approx 800.n^3$	$O(n^3)$	$\approx 262.10^6.n^3$

The algorithm ALS solves each test problem in less than a minute on a Pentium III computer, and the D&C method needs for some test problems about a week on the same computer.

At Step 1, Step 2, Step 4, Step 6, Step 8, and Step 9 the presented heuristic algorithm calculates correspondingly: m times, one time, kn times, $k'n$ times, $n\times(n-1)/2$ and kn times a $n\times n$ determinant. Hence $(2k+k')n + n\times(n+1)/2 +1$ deterninants (with size $n\times n$) will be solved. To calculate one such determinant $2n(n-1)$ multiplications and $2n$ additions (subtractions) are performed. At Step 10 one inverse $n\times n$ matrix is calculated, so that Step 10 costs $O(n^3)$ arithmetical operations It follows that the heuristic algorithms of this type perform $O(\alpha n^3 + \beta mn^2)$ arithmetical operations. The values of α and β depend on the choice of k and k'. If $\alpha \leq m$ and $\beta \leq m$, then the worst-case performance of ALS-type algorithms will cost $O(mn^3 + m^2n^2)$ arithmetical operations. For comparison the computational complexity of FCS algorithm (see [12]) is $O(n^5+mn^3 + m^2n)$ arithmetical operations. The computational complexity of the D&C method [6] depends exponentially on n and increases very rapidly with the increase of the overlapping of both observation data sets.

6. Illustrative example

The following test example with two variables (n=2) and ten training sample observations, with 5 observations in each group ($m_1 = m_2 = 5$), illustrates the performance of the algorithm from Section 3

$$\min z = \sum_{i=1}^{10} \delta_i$$

subject to:
$$\begin{aligned}
-4w_1 - 7w_2 + w_0 + M\delta_1 &\geq 0\\
-5.5w_1 - 5w_2 + w_0 + M\delta_2 &\geq 0\\
-6.5w_1 - 6w_2 + w_0 + M\delta_3 &\geq 0\\
-7w_1 - 10w_2 + w_0 + M\delta_4 &\geq 0\\
-8w_1 - 8w_2 + w_0 + M\delta_5 &\geq 0\\
5.5w_1 + 8w_2 - w_0 + M\delta_6 &\geq \varepsilon\\
7.5w_1 + 4w_2 - w_0 + M\delta_7 &\geq \varepsilon\\
8.5w_1 + 7w_2 - w_0 + M\delta_8 &\geq \varepsilon\\
9.5w_1 + 6w_2 - w_0 + M\delta_9 &\geq \varepsilon\\
10w_1 + 9w_2 - w_0 + M\delta_{10} &\geq \varepsilon,
\end{aligned}$$

where M=10 000 and $\varepsilon = 0.021$.

Solving (5) and (6) is obtained $O_1 = (6.2 , 7.2)$ and $O_1 = (8.2 , 6.8)$.

From (7) $h = (2, -0.4)$.

At Step 1 the initial hyperplane H_0 passes through observation 3 and has the following coefficients:

$w_1 = 2$, $w_2 = -0.4$ and $w_0 = 12.8$.

At Step 2 the hyperplane H_1, passing trough observation 2 and 3 has the following coefficients:

$w_1 = 1$, $w_2 = -1$ and $w_0 = 0.5$.

For this hyperplane $z = 1$, because the sixth inequality is violated.

At Step 4 the set P includes 8 indices of observations: P={1, 4, 5, 6, 7, 8, 9, 10}. Three hyperplanes having $z = 1$ (because $\delta_6 = 1$) are obtained, correspondingly passing through observations: (2), (3), (2), (10) and (3), (5). Their coefficients are:

1) $w_1 = 1$, $w_2 = -1$ and $w_0 = 0.5$ or $\mathbf{w_1 - w_2 = 0.5}$;
2) $w_1 = 4$, $w_2 = -4.5$ and $w_0 = -0.5$ or $w_1 - 4.5w_2 = -0.5$;
3) $w_1 = 1$, $w_2 = -0.75$ and $w_0 = 2$ or $\mathbf{w_1 - 0.75w_2 = 2}$.

Calculating the rating of all variables the following result is obtained: rate(2)=rate(3) = 1; rate(5) = rate(10) = 1.5; rate(4) = rate(8) = 2.5; rate(1) = rate(6) = rate(7) = 3; rate(9) = 4.5. The observations with indices 2, 3, 5, 10 have been included at least one time in the set S_H. Then the set P' = {4, 8}is created at Step 5.

At Step 6 a new hyperplane having $z = 1$ (because $\delta_6 = 1$) is obtained: (8,10). Its coefficients are:

4) $w_1 = 2$, $w_2 = -1.5$ and $w_0 = 6.5$ or $2w_1 - 1.5w_2 = 6.5$.

At Step 8 no better solution has been found.

At Step 9 performing an enumeration cycle arround the hyperplane (3), (5) a new hyperplane having $z = 1$ (because $\delta_6 = 1$) is obtained (5), (7) with coefficients:

5) $w_1 = 8$, $w_2 = -1$ and $w_0 = 56$ or $8w_1 - w_2 = 56$.

Repeating this step one more new hyperplane having $z = 1$ (because $\delta_6 = 1$) is obtained (7), (8) with coefficients:

6) $w_1 = 3$, $w_2 = -1$ and $w_0 = 18.5$ or $3w_1 - w_2 = 18.5$.

At Step 10 the hyperplane through (2), (10) is turned slightly, so that $\bar{a}$ = [0.21, –0.231]. The coefficients of the turned hyperplane are:

2') $w_1 = 4.21$, $w_2 = -4.731$ and $w_0 = -0.5$ or $\mathbf{4.21w_1 - 4.731w_2 = -0.5}$.

The same operation is performed with the hyperplane passing through (5), (7), so that $\bar{a}$=[0.006, –0.006]. The coefficients of the turned hyperplane are:

5') $w_1 = 8.006$, $w_2 = -1.006$ and $w_0 = 56$ or $\mathbf{8.006w_1 - 1.006w_2 = 56}$.

The cut-off value of the hyperplane through (8), (10) is slightly changed:

4') $\hat{w}_0 = w_0 - \varepsilon = 6.5 - 0.021 = 6.479$ or $\mathbf{2w_1 - 1.5w_2 = 6.479}$.

The same operation is performed with the hyperplane through (7, 8):

6') $\hat{w}_0 = w_0 - \varepsilon = 18.5 - 0.021 = 18.479$ or $\mathbf{3w_1 - w_2 = 18.479}$.

Hence here are obtained six “best” hyperplanes having z-value equal to 1, because $\delta_6 = 1$.

7. Conclusions

The paper demonstrates a heuristic approach, giving a simple way to construct polynomial-time algorithms for the two-group classification problem, formulated as a mixed-integer programming problem, which belongs to the class of NP-hard optimization problems.

The ALS algorithm based on the proposed heuristic approach produces near optimal solutions (20-30% among them are equal to the optimal solution) with drastically small computational efforts in comparison to the exact D&C algorithm. These near optimal solutions may be used as initial solutions for the run of exact algorithms.

The obtained very encouraging results by the presented ALS algorithm and the presented approach are good reason to continue the research in this area and to increase further the size of the accessible two-group classification problems.

Acknowledgment. The author would like to thank to professor Ognyan Asparoukhov and Stefan Dantchev (Central Laboratory for Bio-Medical Engineering, Sofia, Bulgaria) and to professor Paul Rubin (Michigan State University, East Lancing, USA), as well as to Dr. Antonio Pedro Duarte Silva (Universidade Catolica Portuguesa, Porto, Portugal), who gave him a lot of test examples and the exe-file performing the "Divide and Conquer" algorithm.

References

1. Abad, P., W. Banks. New LP Based Heuristics for the Classification Problem. – European J. of Oper. Res., **67**, 1993, 88-100.
2. Asparoukhov, O. K., A. Stam. Mathematical Programming Formulations for Two Group Classification with Binary Variables. – Ann. Oper. Res., **74**, 1997, 89-112.
3. Bajgier, S. M. A., Hill. An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminatt Problem. – Decision Sciences, **13**, 1982, 604-618.
4. Bouveyron, C., S. Girard, C. Schmid. Class-Specific Subspace Discriminant Analysis for High-Dimensional Data. – In: Lecture Notes in Computer Science, volume 3940 (C. Saunders et al., Eds.). Berlin Heidelberg, Springer-Verlag, 2006, 139-150.
5. Crawley, D. R. Logistic Discrimination as an Alternative to Fisher's Linear Discriminant Function. – New Zealand Statistics, **14 (2)**, 1979, 21-25.
6. Duarte, Silva, A., A. Stam. A Mixed-Integer Programming Algorithm for Minimizing the Training Sample Misclassification Cost in Two-Group Classification. – Ann. Oper. Res., **74**, 1997, 129-157.
7. Fisher, R. A. The Use of Multiple Measurements in Taxonomy Problems. – Annals of Eugenics, **7**, 1936, 179-188.
8. Freed, N., F. Glover. A Linear Programming Approach to the Discriminant Problem. – Decision Sciences, **12**, 1981, 68-74.
9. Freed, N., F. Glover. Simple But Powerful Goal Formulations for the Discriminant Problem. – European J. of Oper. Res., **7**, 1981, 44-60.
10. Garey, M. R., D. S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco, W. H. Freeman, 1979.
11. Glover, F., S. Keene, B. Duea. A New Class Of Models For The Discriminant Problem. – Decision Sciences, **19**, 1988, 269-280.
12. Gouljashki, V., O. Asparoukhov, S. Danchev. A Heuristic Algorithm for Solving the Classical Two-Group Classification Problem. – In: Proc. of Bulgarian-Russian Seminar "Methods and Algorithms for Distributed Information Systems Design. Theory and Applications". Prof. V. M. Vishnevsky, Dr. Sc. H. Daskalova, Eds. Sofia, 1997, 107-125.
13. Gouljashki, V., O. Asparoukhov. A Heuristic Procedure for a Two-Group Classification Problem. – Problems of Engineering Cybernetics and Robotics, **48**, 1999, 45-52.
14. Guliashki, V. Solving the Mixed-Integer Problem for Classification in Two Groups by Means of Heuristics. EIST 2001. – In: Proc. of the International Scientific Conference on "Energy and Information Systems Technologies". Prof. Cvetko Mitrovski, Assoc. Prof. Rumen Arnaudov, Eds. Vol. III, Bitola, Macedonia, 2001, 808-812.
15. Joachimsthaler, E. A., A. Stam. Four Approaches to the Classification Problem in Discriminant Analysis: An Experimental Study. – Decision Sciences, **19**, 1988, 322-333.
16. Kim, H., H. Park. Relationships between Support Vector Classifiers and Generalized Linear Discriminant Analysis on Support Vectors. University of Minnesota – Computer Science and Engineering Technical Report 004-005, 2004.
17. **http://www.cs.umn.edu/research/technical_reports.php/technical_reports.php?page=report&report_ id=04-005**, (last modified on May 19, 2006).
18. Koehler, G. J., S. S. Erenguc. Minimizing Misclassifications in Linear Discriminant Analysis. – Decision Sciences, **21**, 1990, 63-85.
19. Marinova, G. I. An Accurate Data-Based Random Vector Generator. – In: Proc. of 8th Seminar "Statistical data analysis". Varna, Bulgaria, 1992, 53-59.

20. M a r k o w s k i, C. A., E. P. M a r k o w s k i. Some Difficulties and Improvements in Applying Linear Programming Formulations to the Discriminant Problem. – Decision Sciences, **16**, 1985, 237-247.
21. M a r k o w s k i, C. A., E. P. M a r k o w s k i. An Experimental Comparison of Several Approaches to the Discriminant Problem with Both Qualitative and Quantitative Variables. – European J. of Oper. Res., **28**, 1987, 74-78.
22. M c L a c h l a n, G. J. Discriminant Analysis and Statistical Pattern Recognition. New York, Wiley, 1992.
23. R u b i n, P. Heuristic Solution Procedures for a Mixed-Integer Programming Discriminant Model. – Managerial and Decision Economics, **11**, 1990, 255-266.
24. S m i t h, C. A. B. Some Examples of Discrimination. – Ann. of Eugenics, **13**, 1947, 272-282.
25. S o l t y s i k, R. C., P. R. Y a r n o l d. The Warmack-Gonzalez Algorithm for Linear Two-Category Multivariate Optimal Discriminant Analysis. – Computers & Operations Research, **21**, 1994, 735-745.
26. W a r m a c k, R. E., R. C. G o n z a l e z. An Algorithm for the Optimal Solution of Linear Inequalities and its Application to Pattern Recognition. – IEEE Transactions on Computers, Vol. **C-22**, 1973, 1065-1075.