БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 51
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 51

София . 2001 . Sofia

Bridging Mathemdtical Programming-Based Two-Group Classification with Bayes Decision Approach

Ognian Asparoukhov[1], Plamen Mateev[2]

[1]*Centre of Biomedical Engineering, 1113 Sofia*
[2]*Institute of Mathematics and Informatics, 1113 Sofia*

I. Introduction

Over the years a considerable amount of literature has been accumulated on classification (discriminant analysis), in various fields, including engineering, medical and social sciences, biology, economics, marketing, finance and management (McLachlan, 1992; Ragsdale and Stam, 1992).

The best developed and earliest [15, 16, 25] and appeared [12, 37, 38, 40] was the statistical theory of classification. We will refer to it as a *traditional classification theory*. This theory is based on the *Bayes decision approach* [4, 10] that seeks to divide the space of observations into mutually exclusive and exhaustive regions such that if one observation falls into the i-th region it will be allocated to i-th class (group). The classification regions are defined by minimization of the expected costs due to a wrong decision.

The research conducted over the past twenty years was aimed at the development and/or application of different mathematical approaches for classification, such as: mathematical programming (MP) – based classification [34], neural networks (NN) [24, 29], support vector (SV) learning [32, 36], genetic algorithms (GA) [9, 14]. We will call all these approaches *nontraditional classification approaches*.

The main reason for the development of a such tremendous variety of classification procedures (classifiers) is that *no superior classifier can be found among them*. The usual practice for a particular application is to try as many as possible classifiers in order to choose the best one. Unfortunately, the most of the researchers and users in the field of classification use very restricted number of classifiers – the scientists working in the field of statistical classification use mainly statistical classifiers while the scientists from nontraditional classification fields use mainly classifiers from their own field and some well known (linear – LDF and quadratic – QDF discriminant function, logistic discrimination – LD, nearest neighbors – kNN) statistical classifiers. Rarely [18, 26, 28] papers on statistical discriminant analysis include comparison with nontraditional classifiers (MP, NN, SV).

How could be explained this lack of serious interest by statisticians, working in the field of classical classification, regarding the nontraditional (e.g. MP, NN, SV, GA) classification approaches? In our opinion the main reason for that is *the non - Bayes decision theoretical approach basis* of the most of the nontraditional classification approaches.

The most popular classifiers are the linear ones. They could be constructed using different classification approaches: statistical [3, 12, 13, 18, 25], operational research [5, 39], mathematical programming [5, 11, 30], genetic algorithms [9], support vector [32] and other approaches. *In this paper we will show that any linear classifier could be considered within the Bayes decision theoretic approach framework*. We will carry out our consideration about the mixed integer programming based linear classifier since this is the only classifier that directly minimizes the number of misclassifications and it is the best classifier if the criterion is minimum error rate on the training set. However we would like to stress again that our consideration is valid for any linear classifier.

II. Mathematical programming based classification

During the last twenty years, a class of non-parametric mathematical programming (MP)-based techniques has attracted considerable research attention [33]. We will focus our attention to the mixed-integer programming (MIP) classification approach since only it minimizes directly the number of misclassifications while all statistical procedures minimize this number indirectly by minimizing the value of the misclassification probability. There are several studies devoted to MIP classification algorithms (e.g. [11, 21, 31]). Unfortunately all known MIP-based classification formulations are NP-hard [2] and there is no hope to obtain fast (polynomial time) algorithms for their solving unless $P = NP$.

A number of studies compared the MIP classification method with the most frequently used statistical methods (LDF, QDF, LD) using either real or simulated data [1, 6, 8, 20, 21, 23, 30, 34, 35]. The conclusions of these studies are not uniformly supportive of the MIP classification method but there is a fair amount of support for the statement that it has classified surprisingly well if the data are highly skewed or outlier-contaminated. Very often it clearly outperforms the above mentioned statistical discriminant methods.

The well known MP formulations are based on the geometrical point of view in respect to the discrimination. They construct hyperplane by minimization of some criterion, based on the values proportional to the Euclidean distances of the points to the hiperplane taking into account their position relative to it. Such an approach belongs to the distance-based discrimination. There are a lot of distance measures, proposed for discriminant purposes. Most of these measures have a probability interpretation (they use the probability density functions, covariance matrix etc.) and their application for discriminant purposes is based on the Bayes *decision-theoretic* approach. The Euclidean distance (this is the Mahalanobis distance in casde of independent variables) has a geometrical sense, but not the probability one. Such an approach is not a Bayes *decision-making* approach. It seems that this non-Bayes approach is one of explanations for the lack of serious interest of the statisticians in MP classification.

The purpose of this paper is to show the connection between the Bayes decision-theoretic approach and MIP-based classification. In Section 3 we consider the classical two-group MIP-based formulation of linear classifier's construction. Section 4 is dedicated on the non-parametric estimation of the multivariate normal distribution parameters based on the minimization of the divergence criterion.

III. MIP-based classification

Let us consider the classical sample-based two-group classification problem: g_1 and g_2 are two distinct groups with prior probabilities q_1 and q_2 ($q_1 + q_2 = 1$); a training set of $n = n_1 + n_2$ samples (n_1 from g_1 and n_2 from g_2) is available described by a k-component vector of variables $\boldsymbol{x}^T = (x_1, \ldots, x_k)$. The aim of the discriminant analysis is build a decision function $f(\boldsymbol{w}, \boldsymbol{x})$ such that $\boldsymbol{x} \in g_1$ if $f(\boldsymbol{w}, \boldsymbol{x}) \geq w_0$, otherwise $\boldsymbol{x} \in g_2$; $\boldsymbol{w} = (w_1, \ldots, w_k)^T$ is a vector of the classifier's parameters and w_0 is a cutoff value. The most frequently used decision function $f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} - w_0$ is linear (this is a hyperplane in the k-dimensional attribute space).

The conventional MIP-formulation of the linear classifier's construction is as follows:

$$(1) \qquad \text{minimize} \quad z = \sum_{x_i \in g_1} \frac{q_1}{n_1} y_i + \sum_{x_i \in g_2} \frac{q_2}{n_2} y_i$$

subject to

$$(2) \qquad \begin{aligned} & x_i^T w + M y_i \geq w_0, \text{ if } x_i \in g_1, \\ & x_i^T w + M y_i < w_0, \text{ if } x_i \in g_1, \end{aligned}$$

where w_j ($j = 0, 1, \ldots, k$) are unrestricted (they correspond to the coefficients of the decision hyperplane); the 0/1 integer variable y_i is equal to 1 if the i-th observation is misclassified, and y_i is equal to 0 otherwise (correct classification), $i = 1, \ldots, n$; M is a sufficiently large positive real number. Obviously the objective function z is equal to the overall misclassification error. This is a problem of the discrete optimization and in general case there is not a unique optimal solution.

Definition. Two optimal solutions of the MIP-based classification problem (1)-(2) will differ from classification point of view if and only if their linear classification functions assign at least one sample to different groups.

Let OS_1 and OS_2 are two optimal solutions of the MIP-based classification problem (1)-(2). Therefore they have one and the same minimal value of the object function (1) and they differ at least in one value of the classification function's parameters w_j, $j = 0, 1, \ldots, k$. These two optimal solutions differ from classification point of view if there exists at least one sample, such that if OS_1 assigns it to say g_1 then OS_2 assigns it to g_2 (or vice versa). The number of the optimal solutions is finite since the training sample set is finite. Let $\boldsymbol{v} = (v_1, \ldots, v_k)^T$ and v_0 are the parameter values of the optimal solution OS. These values allow us to assign every observation from the training set to one of the two groups, or in other words we have the following consistent system from n inequalities:

$$(3) \qquad \begin{aligned} & x_i^T w \geq w_0, \text{ if OS assigns } x_i \text{ into } g_1 \text{ (obviously it is possible } x_i \notin g_1), \\ & x_i^T w < w_0, \text{ if OS assigns } x_i \text{ into } g_2 \text{ (obviously it is possible } x_i \notin g_2), \end{aligned} \quad i = 1, \ldots, n,$$

and OS ($w_j = v_j$, $j = 0, 1, \ldots, k$) is a possible solution of the system (3). Obviously every optimal solution OS from classification point of view is a set of infinite number of optimal solutions from mathematical point of view that differ in some values of the classification function parameters (w_j, $j = 0, 1, \ldots, k$), but all these solutions assign each training set observation to one and the same group.

IV. Estimation of the parameters of the multivariate normal distribution based on MIP classification and divergence criterion

We will consider a particular optimal (from classification point of view) solution OS of the classification problem.

Let us accept that the both groups have multivariate normal distribution with common covariance matrix – $N(\mu_i, \Sigma)$, $i = 1, 2$. We look for such estimations of the

means $\tilde{\mu}_i$, $i = 1, 2$ and the covariance matrix $\tilde{\Sigma}$, so that:

a) following the *Bayes approach* (with $\boldsymbol{N}(\tilde{\mu}_i, \tilde{\Sigma})$, $i = 1, 2$) the same assignment of the samples from the training set will be obtained as with the OS (see (3));

b) the divergence criterion between the multivariate normal distributions $\boldsymbol{N}(\tilde{\mu}_i, \tilde{\Sigma})$ and $\boldsymbol{N}(\hat{\mu}_i, \hat{\Sigma})$, $i = 1, 2$, has to be minimum, where $\hat{\mu}_i$, $i = 1, 2$, and $\hat{\Sigma}$ are the *maximum-likelihood* estimations of the means and the covariance matrix.

This is a typical optimization task. First of all let us consider the criterion that should be minimized. The divergence between the multivariate normal distributions $\boldsymbol{N}(\tilde{\mu}_i, \tilde{\Sigma})$ and $\boldsymbol{N}(\hat{\mu}_i, \hat{\Sigma})$ with density functions $\tilde{p}(x/g_i)$ and $\hat{p}(x/g_i)$ is as follows [22]:

$$(4) \qquad D(\boldsymbol{N}(\tilde{\mu}_i, \tilde{\Sigma}) \,||\, \boldsymbol{N}(\hat{\mu}_i, \hat{\Sigma})) = E\{-\ln \frac{\tilde{p}(x/g_i)}{\hat{p}(x/g_i)} / g_i\} =$$

$$= \frac{1}{2} \ln \frac{|\hat{\Sigma}|}{|\tilde{\Sigma}|} + \frac{1}{2} \mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1} - I) + \frac{1}{2} (\tilde{\mu}_i - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (\tilde{\mu}_i - \hat{\mu}_i), \; i = 1, 2.$$

Therefore we should minimize the following criterion:

$$(5) \qquad Q = \sum_{i=1}^{2} n_i D(\boldsymbol{N}(\tilde{\mu}_i, \tilde{\Sigma}) \,||\, \boldsymbol{N}(\hat{\mu}_i, \hat{\Sigma})).$$

Now let us consider the conditions under which the criterion should be minimized.

The class of the optimal (Bayes) classification rules for groups with multivariate normal distribution with common covariance matrix is based on the linear discriminant function (LDF – see [4]):

$$(6) \qquad f_i(x) = w_i^T x + w_{i0}, \quad w_i = \tilde{\Sigma}^{-1} \tilde{\mu}_i, \; w_{i0} = -\frac{1}{2} \tilde{\mu}_i^T \tilde{\Sigma}^{-1} \tilde{\mu}_i + \ln(q_i), \; i = 1, 2.$$

Then the linear classifier has the form: $x \in g_1$ if $f(w, x) = w^T x \geq w_0$ otherwise $\mathbf{x} \in g_2$, where:

$$(7) \qquad w = w_1 - w_2 = \tilde{\Sigma}^{-1} (\tilde{\mu}_1 - \tilde{\mu}_2),$$

$$(8) \quad w_0 = w_{20} - w_{10} = \frac{1}{2} (\tilde{\mu}_1 + \tilde{\mu}_2)^T \tilde{\Sigma}^{-1} (\tilde{\mu}_1 - \tilde{\mu}_2) + \ln\left(\frac{q_2}{q_1}\right) = \frac{1}{2} (\tilde{\mu}_1 - \tilde{\mu}_2)^T w + \ln\left(\frac{q_2}{q_1}\right).$$

Let us now summarize the optimization problem:

$$(9) \quad Q = n[\ln(|\hat{\Sigma}|) - \ln(|\tilde{\Sigma}|) + \mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1} - I)] + n_1 (\tilde{\mu}_1 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (\tilde{\mu}_1 - \hat{\mu}_1) + n_2 (\tilde{\mu}_2 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\tilde{\mu}_2 - \hat{\mu}_2)$$

subject to:

$$(10) \qquad \begin{array}{l} x_i^T w \geq w_0, \text{ if OS assigns } x_i \text{ into } g_1, \\ x_i^T w < w_0, \text{ if OS assigns } x_i \text{ into } g_2, \end{array} \quad i = 1, \ldots, n,$$

$$(11) \qquad w = \tilde{\Sigma}^{-1} (\tilde{\mu}_1 - \tilde{\mu}_2),$$

$$(12) \qquad w_0 = \frac{1}{2} (\tilde{\mu}_1 + \tilde{\mu}_2)^T w + \ln\left(\frac{q_2}{q_1}\right).$$

where w_0 and the components of w, $\tilde{\mu}_1$, $\tilde{\mu}_2$, and $\tilde{\Sigma}$ are unrestricted; q_1, q_2 and all elements of $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\Sigma}$ are real constants.

The considered optimization problem is too complicated and we should simplify it.

IV.1. $\tilde{\mu}_1 = \hat{\mu}_1$ and $\tilde{\mu}_2 = \hat{\mu}_2$

Under the above assumptions the optimization problem is transformed as follows:

(13) minimize $Q = -\ln|\tilde{\Sigma}| + \mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1})$

subject to:

(14) $x_i^T w \gtrless w_0$, if OS assigns x_i into g_1, / $x_i^T w < w_0$, if OS assigns x_i into g_2, $\quad i = 1, \ldots, n$,

(15) $w = \tilde{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$,

(16) $$w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T w + \ln\left(\frac{q_2}{q_1}\right).$$

We will consider two assumptions about the covariance matrix.

IV.2. $\tilde{\Sigma} = \Lambda\hat{\Sigma}\Lambda$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$, $\lambda_i > 0 \ \forall\, i = 1, \ldots, k$.

In this case we have k unknown parameters.
Let $\hat{\Sigma} = \|c_{ij}\|_{i,j=1}^{k}$ and the inverse matrix to exist and $\hat{\Sigma}^{-1} = \|r_{ij}\|_{i,j=1}^{k}$.
Then obviously:

$$\Lambda^{-1} = \mathrm{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_k^{-1}),\ |\Lambda| = \lambda_1\lambda_2\ldots\lambda_k,\ |\Lambda^{-1}| = \lambda_1^{-1}\lambda_2^{-1}\ldots\lambda_k^{-1}$$

$$|\tilde{\Sigma}| = |\Lambda||\hat{\Sigma}||\Lambda| = (\lambda_1\lambda_2\ldots\lambda_k)^2|\hat{\Sigma}| \text{ and } \ln(|\tilde{\Sigma}|) = 2\ln(\lambda_1\lambda_2\ldots\lambda_k) + \ln(|\hat{\Sigma}|)$$

$$\tilde{\Sigma} = \Lambda\hat{\Sigma}\Lambda = \|\lambda_i\lambda_j c_{ij}\|_{i,j=1}^{k} \text{ and } \tilde{\Sigma}^{-1} = (\Lambda\hat{\Sigma}\Lambda)^{-1} = \Lambda^{-1}\hat{\Sigma}^{-1}\Lambda^{-1} = \left\|\frac{c_{ij}}{\lambda_i\lambda_j}\right\|_{i,j=1}^{k}$$

$$\mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1}) = \mathrm{tr}(\Lambda\hat{\Sigma}\Lambda\hat{\Sigma}^{-1}) = \sum_{i=1}^{k}\sum_{j=i}^{k} a_{ij}\lambda_i\lambda_j .$$

Let us denote the components of the k-dimensional vectors as follows:
$\hat{\mu}_1 = (\hat{\mu}_{11}, \hat{\mu}_{12}, \ldots, \hat{\mu}_{1k})^T$, $\hat{\mu}_1 = (\hat{\mu}_{21}, \hat{\mu}_{22}, \ldots, \hat{\mu}_{2k})^T$, $w = (w_1, w_2, \ldots, w_k)^T$.
Then the optimization problem is transformed as follows:

Formulation I.

(17) Minimize $Q = -\ln(|\tilde{\Sigma}|) + \mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1}) = -2\ln(\lambda_1\lambda_2\ldots\lambda_k) + \sum_{i=1}^{k}\sum_{j=i}^{k} a_{ij}\lambda_i\lambda_j$.

subject to:

(18) $x_i^T w \gtrless w_0$ if OS assigns x_i into g_1, / $x_i^T w < w_0$ if OS assigns x_i into g_2, $\quad i = 1, \ldots, n$ (n linear constrains),

(19) $$w_i = \sum_{i=1}^{k}\frac{r_{ij}}{\lambda_i\lambda_j}(\hat{\mu}_1 - \hat{\mu}_2),\ i = 1, \ldots, k \text{ (k fractional non-linear equations)},$$

(20) $$w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T w + \ln\left(\frac{q_2}{q_1}\right).$$

$$\begin{matrix}\lambda_i \geq 1-\varepsilon \\ \lambda_i \leq 1-\varepsilon\end{matrix} \qquad i=1,\ \ldots,\ k \text{ (2}k\text{ linear constrains)}, \tag{21}$$

where $\varepsilon > 0$, q_1, q_2, r_{ij}, a_{ij}, $\hat{\mu}_{1j}$, $\hat{\mu}_{2j}$ ($i, j = 1, \ldots, k$) are real constants.

We include the $2k$ linear constrains (21) since we would like the covariance matrix $\tilde{\Sigma}$ to be close ($\varepsilon > 0$ is a small real number) to the *maximum-likelihood* estimation of the covariance matrix $\tilde{\Sigma}$. (17) is a non-linear function in respect to the unknown $\lambda_1, \lambda_2, \ldots, \lambda_k$. (18), (20), (21) are $n+2k+1$ linear constrains and (19) is a system of k fractional non-linear equations. Therefore the optimization problem (17)-(21) is a problem of fractional non-linear optimization and could be solved by conventional software packages.

As a result of the proposed nonparametric estimation of the covariance matrix ($\tilde{\Sigma} = \Lambda\hat{\Sigma}\Lambda$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$, $\lambda_i > 0\ \forall\ i = 1, \ldots, k$) and the optimization task (17)-(21) we found two multivariate normal distributions $\boldsymbol{N}(\hat{\mu}_i, \tilde{\Sigma})$, $i = 1, \ldots, 2$ ($\hat{\mu}_i$ are the *maximum-likelihood* estimations of the group means), that in accordance with the *Bayes discriminant rule* will allocate one observation to the group with greatest posterior probability and this allocation will coincide with the allocation of the conventional MIP-formulation of the linear classifier's construction (1)-(2).

IV.3. $\tilde{\Sigma} = \lambda\hat{\Sigma}_1 + (1-\lambda)\hat{\Sigma}_2$ where $\lambda \in (0, 1)$ and $\hat{\Sigma}_1$ ($\hat{\Sigma}_2$) is the maximum-likelihood estimation of the first (second) group. In this case we have one unknown parameter.

Let $\hat{\Sigma}_1 = \| c_{ij}^{(1)} \|_{i,j=1}^{k}$ and $\hat{\Sigma}_2 = \| c_{ij}^{(2)} \|_{i,j=1}^{k}$.

Then obviously:

$|\tilde{\Sigma}| = \det_k(\lambda)$ is a k degree polynomial of λ,

$$\tilde{\Sigma}^{-1} = [\lambda\hat{\Sigma}_1 + (1-\lambda)\hat{\Sigma}_2]^{-1} = \| (c_{ij}^1 - c_{ij}^2)\lambda + c_{ij}^2 \|_{i,j=1}^{k\ -1} = \left\| \frac{R_{i,j}^{k}(\lambda)}{\det_k(\lambda)} \right\|_{i,j=1}^{k},$$

where $R_{i,j}^{k}(\lambda)$ – k degree polynomial of λ,

$\mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1}) = \mathrm{tr}(\lambda\hat{\Sigma}_1 + (1-\lambda)\hat{\Sigma}_2\hat{\Sigma}^{-1}) = \lambda\,\mathrm{tr}(\hat{\Sigma}_1 - \hat{\Sigma}_2\hat{\Sigma}^{-1}) + \mathrm{tr}(\hat{\Sigma}_2\hat{\Sigma}^{-1})$ – linear function of λ.

Then the optimization problem is transformed as follows.

Formulation II.

$$\text{minimize } Q = -\ln(|\tilde{\Sigma}|) + \mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1}) = -\ln[\det_k(\lambda)] + \lambda\,\mathrm{tr}(\hat{\Sigma}_1 - \hat{\Sigma}_2\hat{\Sigma}^{-1}) \tag{22}$$

subject to:

$$\begin{matrix} x_i^{\mathrm{T}} w \geq w_0\ , \text{ if OS assigns } x_i \text{ into } g_1, \\ x_i^{\mathrm{T}} w < w_0\ , \text{ if OS assigns } x_i \text{ into } g_2, \end{matrix} \qquad i = 1, \ldots, n, \text{ (}n\text{ linear constrains)}, \tag{23}$$

$$w_i = \frac{1}{\det_k(\lambda)} \sum_{j=1}^{k} R_{ij}^{k}(\lambda)(\hat{\mu}_{1j} - \hat{\mu}_{1j}),\ i = 1, \ldots, k \text{ (}k\text{ fractional non-linear equations)}, \tag{24}$$

$$w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^{\mathrm{T}} w + \ln\left(\frac{q_2}{q_1}\right), \tag{25}$$

where the only variable $\lambda \in (0, 1)$.

The object (criterion) (22) is a non-linear function in respect to the unknown λ. (23), (25) are $n+1$ linear constrains and (24) is a system of k fractional non-linear

equations. Therefore the optimization problem (22)-(25) is a problem of fractional non-linear optimization and could be solved by conventional software packages.

As a result of the proposed nonparametric estimation of the covariance matrix ($\tilde{\Sigma} = \lambda\hat{\Sigma}_1 + (1-\lambda)\hat{\Sigma}_2$ where $\lambda \in (0, 1)$) and the optimization task (22)-(25) we find two multivariate normal distributions $\boldsymbol{N}(\hat{\mu}_i, \tilde{\Sigma})$, $i = 1, 2$ ($\hat{\mu}_i$ are the *maximum-likelihood* estimations of the group means), that in accordance with the *Bayes discriminant rule* will allocate one observation to the group with greatest posterior probability and this allocation will coincide with the allocation of the conventional MIP-formulation of the linear classifier's construction (1)-(2).

In fact **maybe** the true (from statistical point of view) assumption should be not $\tilde{\Sigma} = \lambda\hat{\Sigma}_1 + (1-\lambda)\hat{\Sigma}_2$ where $\lambda \in (0, 1)$. The true assumption [29, p. 106] has to be

$$\tilde{\Sigma} = \frac{\lambda n_1\hat{\Sigma}_1 + (1-\lambda)n_2\hat{\Sigma}_2}{\lambda n_1 + (1-\lambda)n_2}, \text{ where } \lambda \in (0, 1),$$

and $\hat{\Sigma}_1$ ($\hat{\Sigma}_2$) is the maximum-likelihood estimation of the first (second) group) - in this case we have one unknown parameter.

Let $\hat{\Sigma}_1 = \|c_{ij}^{(1)}\|_{i,j=1}^{k}$ and $\hat{\Sigma}_2 = \|c_{ij}^{(2)}\|_{i,j=1}^{k}$.

Then obviously:

$$|\tilde{\Sigma}| = \frac{\det_k(\lambda)}{\lambda n_1 + (1-\lambda)n_2}, \text{ where } \det_k(\lambda) \text{ is a } k \text{ degree polynomial in } \lambda;$$

$$\tilde{\Sigma}^{-1} = \left(\frac{\lambda n_1\hat{\Sigma}_1 + (1-\lambda)n_2\hat{\Sigma}_2}{\lambda n_1 + (1-\lambda)n_2}\right) = [\lambda n_1 + (1-\lambda)n_2]\,\|(n_1c_{ij}^1 - c_{ij}^2)\lambda + n_1c_{ij}^2\|_{i,j=1}^{k\ -1} =$$

$$= [\lambda n_1 + (1-\lambda)n_2]\left\|\frac{R_{i,j}^{k}(\lambda)}{\det_k(\lambda)}\right\|_{i,j=1}^{k}, \text{ where } R_{i,j}^{k}(\lambda) - k \text{ degree polynomial of } \lambda;$$

$$\mathrm{tr}(\tilde{\Sigma}\hat{\Sigma}^{-1}) = \mathrm{tr}\left(\frac{\lambda n_1\hat{\Sigma}_1 + (1-\lambda)n_2\hat{\Sigma}_2}{\lambda n_1 + (1-\lambda)n_2}\,\hat{\Sigma}^{-1}\right) = \frac{\lambda}{\lambda n_1 + (1-\lambda)n_2}\,\mathrm{tr}(n_1\hat{\Sigma}_1\hat{\Sigma}^{-1} - n_2\hat{\Sigma}_2\hat{\Sigma}^{-1}) +$$

$$+ \frac{n_2}{\lambda n_1 + (1-\lambda)n_2}\,\mathrm{tr}(\hat{\Sigma}_2\hat{\Sigma}^{-1}) - \text{linear function of } \lambda.$$

Then the optimization problem is transformed as follows:

Formulation IIa:

Is the same as (22)-(25) substituting corresponding estimates.

V. Recursive quadratic programming formulation (Formulation III)

Our aim here is to create a formulation with quadratic object, all constrains of which being linear. Let us consider again the equation (12)

$$w_0 = \frac{1}{2}(\tilde{\mu}_1 + \tilde{\mu}_2)^t w + \ln\left(\frac{q_2}{q_1}\right),$$

Let us assume that $\tilde{\mu}_1 + \tilde{\mu}_2 = \hat{\mu}_1 + \hat{\mu}_2$.

Then (12) is transformed into the following linear constrain:

$$w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^t w + \ln\left(\frac{q_2}{q_1}\right).$$

Step 0. Let $\tilde{\mu}_i(0) = \hat{\mu}_i$, $i = 1, 2$.

Step 1. Calculate the covariance matrix $\tilde{\Sigma}(1)$ of the training sample with means $\tilde{\mu}_i = \tilde{\mu}_i(0) = \hat{\mu}_i$, $i = 1, 2$, using the conventional maximum-likelihood estimation:

$$(26)\quad \tilde{\Sigma}(1) = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{x_j \in g_i}(x_j - \tilde{\mu}_i(0))(x_j - \tilde{\mu}_i(0))^T = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{x_j \in g_i}(x_j - \tilde{\mu}_i)(x_j - \tilde{\mu}_i)^T = \hat{\Sigma}.$$

Under the above assumptions the optimization problem (9)–(12) is transformed as follows:

$$(27)\quad \text{minimize } Q(\text{step1}) = \sum_{i=1}^{2} n_i D(N(\tilde{\mu}_i, \tilde{\Sigma}(1)) \| N(\hat{\mu}_i, \hat{\Sigma})) = \sum_{i=1}^{2} n_i D(N(\tilde{\mu}_i, \hat{\Sigma}) \| N(\hat{\mu}_i, \hat{\Sigma})) =$$

$$= n_1(\tilde{\mu}_1 - \hat{\mu}_1)^t \hat{\Sigma}^{-1}(\tilde{\mu}_1 - \hat{\mu}_1) + n_2(\tilde{\mu}_2 - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\tilde{\mu}_2 - \hat{\mu}_2)$$

subject to

$$(28)\quad \begin{array}{ll} x_i^T w \geq w_0 & \text{if OS assigns } x_i \text{ into } g_1, \\ x_i^T w < w_0 & \text{if OS assigns } x_i \text{ into } g_2, \end{array} \qquad i = 1, \ldots, n,$$

$$(29)\quad w = \hat{\Sigma}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2),$$

$$(30)\quad w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T w + \ln\left(\frac{q_2}{q_1}\right),$$

$$(31)\quad \tilde{\mu}_1 + \tilde{\mu}_2 = \hat{\mu}_1 + \hat{\mu}_2,$$

where w_0 and the components of w, $\tilde{\mu}_1$, $\tilde{\mu}_2$ are unrestricted; q_1, q_2 and all elements of $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\Sigma}$ are real constants.

The object (criterion) (27) is a quadratic function in respect to the unknown $\tilde{\mu}_{ij}$ ($i = 1, 2$, $j = 1, \ldots, k$), (28) is a system of n linear inequalities, (29) and (31) are two systems of k linear equations and (30) is a linear equation. Therefore we have quadratic object (27) and ($n+2k+1$) linear constrains (28)–(31), or the described optimization problem is a task of a quadratic mathematical programming.

The result of step 1 is as follows:

$$(32)\quad (\tilde{\mu}_1(1), \tilde{\mu}_2(1)) = \operatorname{argmin} Q(\text{step1}), \text{ where } Q_1 = \min Q(\text{step1}) =$$

$$= n_1(\tilde{\mu}_1(1) - \hat{\mu}_1)^T \hat{\Sigma}^{-1}(\tilde{\mu}_1(1) - \hat{\mu}_1) + n_2(\tilde{\mu}_2(2) - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\tilde{\mu}_2(2) - \hat{\mu}_2).$$

Step s. Calculate the covariance matrix $\tilde{\Sigma}(s)$ of the training sample with means $\tilde{\mu}_i = \tilde{\mu}_i(s-1)$, $i = 1, 2$, using the conventional maximum-likelihood estimation:

$$(33)\quad \tilde{\Sigma}(s) = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{x_j \in g_i}(x_j - \tilde{\mu}_i(s-1))(x_j - \tilde{\mu}_i(s-1)).$$

Under the above assumptions the optimization problem (9)–(12) is transformed as follows:

$$(34)\quad \text{minimize } Q(\text{step } s) = \sum_{i=1}^{2} n_i D(N(\tilde{\mu}_i, \tilde{\Sigma}(s)) \| N(\hat{\mu}_i, \hat{\Sigma})) = n \ln|\hat{\Sigma}| - n \ln|\tilde{\Sigma}(s)| +$$

$$+ n \operatorname{tr}(\tilde{\Sigma}(s)\hat{\Sigma}^{-1} - 1) + n_1(\tilde{\mu}_1 - \hat{\mu}_1)^t \hat{\Sigma}^{-1}(\tilde{\mu}_1 - \hat{\mu}_1) + n_2(\tilde{\mu}_2 - \hat{\mu}_2)^t \hat{\Sigma}^{-1}(\tilde{\mu}_2 - \hat{\mu}_2)$$

subject to

$$x_i^T w \geq w_0, \text{ if OS assigns } x_i \text{ into } g_1, \quad i=1, \ldots, n, \tag{35}$$
$$x_i^T w < w_0, \text{ if OS assigns } x_i \text{ into } g_2,$$

$$w = \tilde{\Sigma}(s)^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2), \tag{36}$$

$$w_0 = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T w + \ln\binom{q_2}{q_1}, \tag{37}$$

$$\tilde{\mu}_1 + \tilde{\mu}_2 = \hat{\mu}_1 + \hat{\mu}_2. \tag{38}$$

In other words the difference between two consecutive steps is the matrix $\tilde{\Sigma}(s)$ which influences the object and the k linear equations (i. e. (44), (51)). Each step is a task of a quadratic mathematical programming – quadratic object and $(n + 2k + 1)$ linear constrains.

The result of step s is as follows:

$$(\tilde{\mu}_1(s), \tilde{\mu}_2(s)) = \underset{\tilde{\mu}_1, \tilde{\mu}_2}{\operatorname{argmin}} Q(\text{step } s), \text{ where} \tag{39}$$

$$Q_s = \min_{\tilde{\mu}_1, \tilde{\mu}_2} Q(\text{step } s) = n\ln|\hat{\Sigma}| - n\ln|\tilde{\Sigma}(s)| + n\operatorname{tr}(\tilde{\Sigma}(s)\hat{\Sigma}^{-1} - 1) +$$
$$+ n_1(\tilde{\mu}_1(s) - \hat{\mu}_1)^t\hat{\Sigma}^{-1}(\tilde{\mu}_1(s) - \hat{\mu}_1) + n_2(\tilde{\mu}_2(s) - \hat{\mu}_2)^t\hat{\Sigma}^{-1}(\tilde{\mu}_2(s) - \hat{\mu}_2).$$

If we prove that

$$Q_1 \geq Q_2 \geq \ldots \geq Q_{s-1} \geq Q_s \geq \ldots, \tag{40}$$

then the result would be that our recursive procedure converges as follows:

$$\tilde{\mu}_i(s),\ i = 1, 2 \xrightarrow{s\to\infty} \tilde{\mu}_i, \text{ so that } \tilde{\Sigma}(s) \xrightarrow{s\to\infty} \tilde{\Sigma}, \tag{41}$$

where $\tilde{\Sigma}$ is a maximum – likelihood estimation of the covariance matrix in respect to the training sample and $\tilde{\mu}_i(s)$, $i = 1, 2$.

For now we have no prove about the above statement, although it seems that it would hold in the practice.

As a result of the proposed non-parametric recursive estimation we find two multivariate normal distributions $N(\tilde{\mu}_i, \tilde{\Sigma})$, $i = 1, 2$, that in accordance with the *Bayes discriminant rule* will allocate one observation to the group with greatest posterior probability and this allocation will coincide with the allocation of the conventional MIP-formulation of the linear classifier's construction (1)-(2).

Let us have more than one solution optimal from the classification point of view. Then we will decide the respective optimization problem (formulations I, II or III) for each of them and will choose the solution with minimum values of the divergence criterion Q.

VI. Conclusion

In this paper we show that any linear classifier could be considered within the Bayes decision theoretic approach framework. We will carry out our consideration about the mixed integer programming based linear classifier since this is the only classifier that directly minimizes the number of misclassification and it is the best classifier if the criterion is the minimum of the training set misclassification error rate. However we would like to stress again that our consideration is valid for any linear classifier.

We consider the two-group mixed integer based classification and show its connection with the Bayes decision theoretic approach. The conventional MIP-formula-

tion of the linear classifier's construction minimizes the overall misclassification error. This is a NP-hard optimization task which result is a set of infinite number of optimal from mathematical point of view solutions that differ in some value(s) of the classification function's parameter(s). All of them assign each training set observation into one and same group. In other words we have a consistent system from n inequalities (n – the number of the training set observations). We assume that the two groups have multivariate normal distribution with common covariance matrix - the class of optimal (Bayes) classification rules in this case is based on the linear discriminant function. Three optimization formulations, based on the minimization of the divergence criterion under given constrains are proposed. Two of the formulations are nonlinear optimization problems, while the last recursive formulation is a task of quadratic mathematical programming (quadratic object and linear constrains). As a result of the proposed non-parametric estimation we find two multivariate normal distributions that in accordance with the *Bayes discriminant rule* will allocate one observation to the group with greatest posterior probability and this allocation will coincide with the allocation of the conventional MIP-formulation of the linear classifier's construction.

References

1. Abad, P. L, W. J. Banks. New LP based heuristics for the classification problem. – European Journal of Operational Research, **67**, 1993, 88-100.
2. Amaldi E., V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. – Theoretical Computer Science, **209**, 1998, 237-260.
3. Anderson, J. A. Separate sample logistic discrimination. – Biometrika, **59**, 1972, 19-35.
4. Anderson, T. W. An Introduction to Multivariate Statistical Analysis. New York, Wiley (2nd ed.), 1984.
5. Asparoukhov, O., P. Rubin. Oscillation heuristics for the two-group classification problem, 2000 (submitted).
6. Asparoukhov, O., A. Stam. Mathematical programming formulations for two group classification with binary variables. Annals of Operations Research, **74**, 1997, 89-112.
7. Asparoukhov, O., S. Danchev S. Discrimination and Classification in the presence of Binary variables. – Biocybernetics and Biomedical Engineering, **17**, 1997, No 1-2, 25-39.
8. Bajgier, S. M., A. Hill. An experimental comparison of statistical and linear programming approaches to the discriminant problems. – Decision Sciences, **13**, 1982, 604-618.
9. Conway, D. G., A. V. Cabot, M. A. Venkataraman. A genetic algorithm for discriminant analysis. – Annals of Operations Research, **78**, 1998, 71-82.
10. Das Gupta, S. Theories and methods in classification: a review. – In: Discriminant Analysis and Applications, T. Cacoullos (Ed.), New York: Academic Press, 1973, 77-137.
11. Duarte Silva, A., A. Stam. A mixed integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. - Annals of Operations Research, **74**, 1997, 129-157.
12. Fisher, R. A. The use of multiple measurements in taxonomic problems. – Annals of Eugenics, **7**, 1936, 179-188.
13. Friedman, J. H., T. Hastie, R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting (with discussion). Annals of Statistics, 1998 (to appear).
14. Goldberg, D. E. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading, MA, 1989.
15. Goldstein, M., W. R. Dillon. Discrete Discriminant Analysis. New York, Wiley, 1978.
16. Hand, D. J. Construction and Assessment of Classification Rules. Chichester, Wiley, 1997.
17. Hastie, T., R. Tibshirani, A. Buja. Flexible discriminant analysis by optimal scoring. – Journal of the American Statistical Association, **89**, 1994, 1255-1270.
18. Hastei, T., R. Tibshirani. Classification by pairwise coupling. – Annals of Statistics, **26**, 1998, No 2, 451-471.
19. Huberty, C. J. Applied Discriminant Analysis. New York: Wiley, 1994.
20. Joachimsthaler, E. A., A. Stam. Mathematical Programming approaches for the classification problem in two-group discriminant analysis. - Multivariate Behavioral Research, **25**, 1990, 427-454.

21. Koehler, G. J., S. S. Erenguc. Minimizing misclassifications in linear discriminant analysis. - Decision Sciences, **21**, 1990, 63-85.
22. Kulback, S. Information theory and statistics. New York, Wiley, 1959.
23. Lam, K. F., E. U. Choo, W. C. Wedley. Linear goal programming in estimation of classification probability. - European Journal of Operational Research, **67**, 1993, 101-110.
24. Lippmann, R. P. Pattern classification using neural networks. - In: IEEE Communication Magazine, 1989, 47-64.
25. McLachlan, G. J. Discriminant Analysis and Statistical Pattern Recognition. New York, Wiley, 1992.
26. Nath, R., W. M. Jackson, T. W. Jones. A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis. - J. Statist. Comput. Simul., **41**, 1992, 73-93.
27. Ragsdale, C. T., A. Stam. Introducing discriminant analysis to the business statistics curriculum. - Decision Sciences, **23**, 1992, 724-745.
28. Ripley, B. D. Neural networks and related methods for classification. - J. R. Statist. Soc. B, **56**, 1994, No 3, 409-456.
29. Ripley, B. D. Pattern Recognition and Neural Networks. Cambridge, Cambridge University Press, 1995.
30. Rubin, P. A. Heuristic solution procedures for a mixed-integer programming discriminant model. - Managerial and Decision Economics, **11**, 1990, 255-266.
31. Rubin, P. Solving mixed integer classification problems by decomposition. - Annals of Operations Research, **74**, 1997, 51-64.
32. Scholkopf, B. Support Vector Learning. Munich, R. Oldenbourgs Verlag, 1997.
33. Stam, A. MP approaches to classification: issues and trends. - Annals of Operations Research, **74**, 1997, 1-36.
34. Stam, A., 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, **46**, 1990, 113-122.
35. Stam, A., D. G. Jones. Classification performance of mathematical programming techniques in discriminant analysis: Result for small and medium sample size. - Managerial and Decision Economics, **11**, 1990, 243-253.
36. Vapnik, V. The Nature of Statistical Learning Theory. New York, Springer Verlag, 1995.
37. Wald, A. Contributions to the theory of statistical estimation and testing hypothesis. - Ann. Math. Statist., **10**, 1939, 299-326.
38. Wald, A. Statistical decision functions. - Ann. Math. Statist., **20**, 1949, 165-205.
39. Warmack, R. E., R. C. Gonzales. An algorithm for the optimal solution of linear inequalities and its application to pattern recognition. - In: IEEE Trans. on Computers, **C22**, 1973, 1065-1075.
40. Welsh, B. L. Note on discriminant functions. - Biometrika, **31**, 1939, 218-220.

Математическое программирование с точки зрения Байесовской теории принятия решений в случае классификации двух классов

Огнян Аспарухов[1], *Пламен Матеев*[2]

[1]*Центр биомедицинской кибернетики, 1113 София*
[2]*Институте математики и информатики, 1113 София*

(Резюме)

Статья относится к области методов классификации, известной как дискриминантный анализ. Авторами сделана попытка показать, что "нетрадиционные методы", такие как метод смешанного целочисленного программирования, могут быть приведены к традиционной Байесовской постановке при известных предположениях о статистико-вероятностной структуре задачи. Вводится определение различных оптимальных решающих функции, основывающееся на виде ошибок. Выводится критерий минимизации и описывается соответствующая итерационная процедура.