БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 48
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 48

София . 1999 . Sofia

A Heuristic Procedure for a Two-Group Classification Problem[1]

Vassil G. Gouljashki, Ognian K. Asparoukhov***

**Institute of Information Technologies, 1113 Sofia*
***Centre of Biomedical Engineering, 1113 Sofia*

1. Introduction

Recently, a class of nonparametric mathematical programming (MP)-based techniques has attracted considerable research attention – linear programming (LP) and mixed-integer programming (MIP) approaches to the discriminant problems (Stam, 1997). The well known MP formulations are based on the geometrical point of view with respect to the discrimination. More precisely they construct hyperplane (linear function of the attributes) by minimization of some criterion, based on the values proportional to the Euclidean distances (perpendicular from every observation to this hyperplane) or/and their signs. Such an approach is part of the distance-based discrimination. A lot of studies have been devoted to the comparison of the most well known MP-based (different variants of the LP and MIP) formulations between them and with the most frequently used statistical methods (LDF – linear discriminant function, QDF – quadratic discriminant function, LR – logistic regression) – in terms of their classification performance, using either real or simulated data (Abad and Banks [1], Asparoukhov and Dan-chev [3], Asparoukhov and Stam [4], Bajgier and Hill [5], Duarte Silva and Stam [6], Joachimsthaler and Stam [7, 8], Koehler and Erenguc [9], Lam, Choo and Wedley [10], Nath, Jackson and Jones [12], Rubin [13, 14], Stam and Joachim-sthaler [16, 17], Stam and Jones [18]). The conclusions of these studies are not uniformly supportive of the MP-based methods but there is a fair amount of support for the statement that the MIP methods have classified surprisingly well if the data are highly skewed or outlier-contaminated. Very often these methods clearly outperform the above mentioned statistical discriminant methods. That is why we will direct our attention to the MIP-based classification procedures.

Let us consider a classical sample-based two-group classification problem: g_1 and g_2 are two distinct groups; there is available a training sample of n (n_1 from g_1 and n_2

[1]This research was supported by the National Science Fund of Bulgaria, grants No I-623/96 and I-515/95.

from g_2) observations whose objects are described by a m-component vector of attributes $\boldsymbol{x}^T=(x_1,\ldots,x_m)$. The aim of the classification analysis is a decision function (classifier) $f(\boldsymbol{w},\boldsymbol{x})$ to be found, such that $\boldsymbol{x}\in g_1$ if $f(\boldsymbol{w},\boldsymbol{x}) \geq w_0$, otherwise $\boldsymbol{x}\in g_2$; $\boldsymbol{w}$ is a vector of the decision function's parameters and w_0 is a cutoff value. The most frequently used decision function is linear – $f(\boldsymbol{w},\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}$ – this is a hyperplane in the m-dimensional attribute space.

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

$$\text{minimize } z=\sum_{i=1}^{n} y_i$$

subject to:

$$\boldsymbol{x}^T\boldsymbol{w} + M y_i \geq 0, \quad i\in g_1; i=1,\ldots,n_1,$$
$$\boldsymbol{x}^T\boldsymbol{w} - M y_i < 0, \quad i \in g_2; i = n_1+1,\ldots, n_1+n_2,$$

where w_k (k=0, 1, ..., m) are unrestricted (they correspond to the coefficients of the decision hyperplane); the 0/1 integer variable y_i=1 if the i-th observation is misclassified, and y_i=0 otherwise (correct classification), i=1, ..., n; M is a sufficiently large positive real number. Obviously the objective function z is equal to the number of misclassifications.

MIP models directly attack the objective of minimizing the number of misclassifications at high computational cost, whereas LP models attack that objective indirectly at lower cost (Bajgier and Hill, [5]). There are several studies devoted to MIP classification algorithms (e.g. Warmack and Gonzalez [20], Soltysik and Yarnold [15], Duarte Silva and Stam [6], Rubin [13], Koehler and Erenguc [9]). Unfortunately all known MIP-based classification formulations are NP-hard (Amaldi and Kann, [2]) and there is no hope to obtain fast (polynomial time) algorithms for their solving unless P=NP. Almost all authors use the "branch and bound" techniques. These techniques are well studied in the general case (for solving arbitrary MIP problem) but they are not efficient, since they have exponential computational complexity and do not take into account the specificity of the classification problem. That is why there is reason to consider the use of heuristic procedures that give competitive accuracy with the MIP models but are substantially faster. Such heuristics can be used also to provide a good initial feasible solution to a branch-and-bound mixed-integer classification models.

The purpose of this paper is to introduce such a heuristic that is based on the fast combinatorial search taking into account the specificity of the classification problem. The paper also presents preliminary test results for the proposed heuristic, the two known heuristics (Rubin [13]), the MIP-based classification algorithm (Duarte Silva and Stam [6]) and the linear discriminant analysis (LDF), using randomly generated data with multivariate normal distribution.

2. Heuristic procedure based on a fast combinatorial search

The proposed heuristic is based on a mathematical statement and a heuristic assumption.

The first is considered and discussed in Warmack and Gonzalez [20]. We will explain briefly this idea. Let us consider the described in the above section classical sample-based two-group discriminant problem: two groups (g_1, g_2), a training set from n (n_1 from g_1, n_2 from g_2) observations described by m attributes x_1, ..., x_m. The aim is the linear classifier (decision hyperplane) with the minimum number of misclassifications to be constructed so that if $\boldsymbol{w}^T\boldsymbol{x}>w_0$ then assign $\boldsymbol{x}$ to g_1; otherwise – assign $\boldsymbol{x}$ to g_2, where $\boldsymbol{w}^T=(w_1, w_2, \ldots, w_m, w_0)$ is a weight vector and w_0 is a cutoff value. Therefore we have

a system of n linear inequalities (one per each observation) and are looking for a minimal system of r inequalities (misclassifications) without which the other $n-r$ linear inequalities (correct classifications) form a consistent system. This is a problem of a discrete optimization and in general case there is not a unique solution. If we replace the inequalities with equalities then each observation forms a hyperplane in the $(m+1)$-dimensional attribute space. Warmack and Gonzalez [20] use the term edge for the intersection of these m observation hyperplanes. Each edge is described by a system of m homogeneous equations (observations) with $m+1$ variables (that means each edge is described by the hyperplane containing these m observations). Their algorithm proceeds by iteration through sequence of edges and exploits actively the following assumption (Haar condition): every $m \times (m+1)$ matrix formed by the intersection of m observation hyperplanes contains at least one $m \times m$ submatrix which is nonsingular. In other words the Haar condition means that is no $m+1$ points (training observations) that lie in a hyperplane in the m-dimensional attribute space and also there is no m points that lie in a $(m-2)$-dimensional plane. It may be accepted that the Warmack – Gonzalez algorithm contains in generally two steps:

– First step: Construction of a hyperplane from combinations of m observations and finding its misclassifications (violated inequalities) accepting that the m observations of the hyperplane are correctly classified. Warmack and Gonzales [20], proved (this is the mathematical statement we used) that the end of this step is an optimal hyperplane (with the minimum number of misclassifications).

– Second step: A linear transformation of the optimal hyperplane, so that the m points (observations) of it go into the subspace corresponding to the group of each of them, while the other $n-m$ points keep the subspace defined by the first step (Warmack and Gonzalez [20], Appendix A).

We will base our heuristic procedure on the above discussed mathematical statement and will search for "promising" (from classification point of view) combinations of m training observations.

The heuristic assumption is that some classifiers hardly depend from the closest (in Euclidean sense) points (training observations). This intuitive assumption is a basis for most nonparametric statistical discriminant procedures as kernel and nearest neighbour estimators, neural networks etc. Obviously the insignificant translation and/or rotation of given hyperplane could change the subspace (in respect to it) mainly of the closest points. Therefore it seems that the closest points are the most "promising" that should be taken into account when we would like to improve the classification accuracy of one decision hyperplane. Or with other words if we use m training observations (the Haar condition holds) to construct the decision hyperplane in the m-dimensional attribute space, then we should replace some of these m observations (points) with the closest (to the hyperplane) points and should investigate the accuracy of the new hyperplanes.

Based on the discussed above mathematical statement and heuristic assumption the main idea of our heuristic procedure is at every step to try to improve the obtained up to now decision hyperplane by forming of new combination of m observations taking:

$m-1$ observations from the set M (we call it the set of the "most promising" observations);

and one observation from the set P (we call it the set of the "promising" observations).

3. Fast combinatorial search (FCS) heuristic algorithm

We will use the following notation during the description of the FCS algorithm:

$x_i=(x_{i1},\ldots,x_{im})'$ – i-th training observation with the values of its attributes;

best – the best (minimum) number of misclassifications;

$M(i)$ – the set of the "most promising" observations of step i;
$P(i)$ – the set of the "promising" observations of step i.

Step 0. Initial solution

Compute the mean vectors O_1 (group g_1) and O_2 (group g_2). Construct the hyperplane H_0 that is orthogonal to the line O_1O_2. The m coefficients of H_0 are: $\boldsymbol{w}=O_2-O_1$. The cutoff value of this hyperplane $w_0=0$. This initial decision hyperplane is equivalent to the linear discriminant function in the presence of independent attributes.

$$best = \min(n_1, n_2);$$

$i=1$. Let $g_i \in H_0$. We will denote this hyperplane by $H_0(\boldsymbol{x}_i)$ – obviously its cutoff value is $w_0=-\boldsymbol{x}_i\boldsymbol{w}^T$. Compute error$(\boldsymbol{x}_i)$ = number of misclassifications for this hyperplane.

If $best >$ error$(\boldsymbol{x}_i)$ then:

$best =$ error $(\boldsymbol{x}_i)$;

$M(0)=\{$the m closest observations to $H_0(\boldsymbol{x}_i)$, including $\boldsymbol{x}_i\}$;

$P(0)=\{$the next k (after the first m) closest observations to $H_0(\boldsymbol{x}_i)\}$;

$i = i + 1$. Do while $i \le n$.

Step 1. Investigation of the combinations between the elements of $M(0)$ and $P(0)$. Construct a hyperplane containing the m observations (points) of $M(0)$. Construct a hyperplane for every combination (of m points) that consists of $m-1$ observations from $M(0)$ and one observation from $P(0)$ – the number of these combinations is $m \times k$. Save the 20 best solutions (hyperplanes). For every one of these 20 hyperplanes create its $M(1)$ and $P(1)$, where $M(1)$ contains the m training observations that form this hyperplane and $P(1)$ contains the v closest misclassification observations (in respect to this hyperplane). Update the value of $best$. Obviously in general more than one hyperplane will have $best$ number of misclassifications and some of these 20 hyperplanes will have number of misclassifications more than $best$. Let us denote the next (by its goodness) error between the 20 hyperplanes with $bestnext$.

Step 2. Investigation of the combinations between the elements of $M(1)$ and $P(1)$ for every hyperplane with error = $best$ or error = $bestnext$. Every time update the best 20 solutions if the current error is less or equal to $best$. For the last obtained solution with error = $best$ create $M(2)$ and $P(2)$ per every attribute. Let us take the j-th attribute $\boldsymbol{x}_j$ and let denote by $O(o_1, \ldots, o_j, \ldots, o_m)$ the middle point of the segment O_1O_2. . Take the $2m$ closest observations to the considered best hyperplane (solution). Assign everyone of these closest observations to the set of:

– "most promising" observations $M(2, j)$ if its j-th attribute is less or equal to o_j;

– "promising" observations $P(2, j)$ if its j-th attribute is greater than o_j.

Step 3. Investigation of the combinations between the elements of $M(2, j)$ and $P(2, j)$, $j=1, \ldots, m$.

Let us accept that card $(M(2, j))$ i card$(P(2, j))$ and let the elements of $M(2, j)$ are $a(1), \ldots, a(s_1)$ and the elements of $P(2, j)$ are $b(1), \ldots, b(s_2)$, $s_1+s_2=2m$, s_1 i s_2. Investigate the following combinations:

$$a(1), \ldots, a(m), b(1); \ldots; a(1), \ldots, a(m), b(s_2);$$

$$a(2), \ldots, a(m+1), b(1); \ldots; a(2), \ldots, a(m+1), b(s_2);$$

$$\ldots$$

$$a(s_1-m+1), \ldots, a(s_1), b(1); \ldots; a(s_1-m+1), \ldots, a(s_1), b(s_2);$$

and update the best solutions if error $\le best$.

Step 4. Only for the last obtained solution with error = *best* repeat Step 1.

Step 5. A linear transformation of the optimal hyperplane.

The aim of this step is its *m* training observations to go into the subspace corresponding to the group of each of them, while the other *n-m* points keep their subspace (Warmack and Gonzalez [20], Appendix A).

4. Experiments and Discussion

To evaluate the efficacy of FCS heuristic algorithm, tests were performed using randomly generated multivariate normal distribution (two groups) as follows: a) 4, 5 and 6 attributes; 100 observations (50 per group) – 20 data sets were generated per every attributes' number; *b*) 4, 5 and 6 attributes; 200 observations (50 per group) – 20 data sets were generated per every attributes's number. For all simulations the common covariance matrix was equal to unity matrix and all components of mean vector per group were equal to 1.0 per group 1 and 0.8 per group 2.

The FCS algorithm has two parameters – *k* (Step 0) and *v* (Step 1) that correspond to the respective number of closest observations (points). These parameters take the following values during our simulated study: *k*=25 and *v*=5.

The experiment includes 20 simulated data sets per every combination of Na and No.

Table 1.

Na	No	Mean number of misclassifications					Mean number of major pivots			
		D&C	FCS	H1	H2	LDF	D&C	FCS	H1	H2
4	100	14.9	16.0	19.8	17.4	20.6	53819	371.6	107.3	1197.1
5	100	11.8	13.1	18.1	14.8	17.3	190730	397.1	141.0	1631.8
6	100	8.9	10.6	15.4	12.0	15.1	387433	455.1	189.9	3164.2
4	200	33.8	35.4	44.3	39.6	41.7	238144	482.5	215.7	2507.2
5	200	26.7	28.9	36.2	31.9	34.5	1245623	551.7	302.7	3980.6
6	200	21.9	24.3	35.3	28.7	30.5	3616615	533.5	341.9	4413.3

Table 1 (continuation)

Na	No	Mean number of minor pivots			Mean pricing			
		D & C	H1	H2	D&C	FCS	H 1	H2
4	100	16544.8	2978.7	29903.4	883152.6	229800	11799.3	133492.3
5	100	47909.9	3590.4	39143.8	3317446.7	521450	15762.0	184315.5
6	100	74610.2	4477.5	68501.8	6877750.0	78750	21571.7	362617.0
4	200	84291.6	11696.5	129029.5	4871555.6	765650	45279.9	529898.6
5	200	343491.7	14779.9	185061.5	25318089.7	955750	64070.2	847418.2
6	200	821557.9	15906.2	200442.8	75362453.8	1170100	73037.7	947413.4

Table 1. Legend:

Na – number of attributes; No – number of observations;
D&C – Divide and Conquer algorithm (Duarte Silva and Stam [6]);
FCS – Fast combinatorial search heuristic;
H1 – Heuristic 1 (Rubin, 1990); H2 – Heuristic 2 (Rubin [13]).

We compare the proposed heuristic algorithm FCS with the following discrimination procedures: a) exact algorithm "Divide and Conquer" (D&C) proposed by Duarte Silva and Stam [6]; b) two heuristic procedures (H1 and H2) developed by Paul Rubin [13]; c) linear discriminant analysis (LDF).

The computational efforts of every algorithm (except LDF) were evaluated by:

a) *Number of major pivots*. A major pivot is one where the selected nonbasic variable is pivoted into the basis (about m arithmetical operations). The FCS algorithm performs one major pivot for a construction of every hyperplane passing through m training observations.

b) *Number of minor pivots*. A minor pivot is one in which the selected nonbasic variable remains nonbasic by going from its upper bound to lower bound or from its lower bound to upper bound (m arithmetical operations). The FCS algorithm is a combinatorial algorithm, based on geometrical interpretations of the classification problem and it does not carry out minor pivots.

c) *Pricing*. "Pricing" refers to the sub-routine of the Simplex method in which this algorithm spends most of its time. A pricing is performed each time a column (associated with a nonbasic variable which is candidate to enter the basis) is evaluated ("priced") regarding to its impact to the objective function. The pricing corresponds to about $2(m+1)$ arithmetical operations.

Table 1 summarizes the statistics for number of misclassifications and computational efforts of investigated classification procedures (algorithms).

The obtained results assure us of the efficacy of the proposed heuristic procedure since its number of misclassifications is:

– definitely less than LDF, H1 and H2 and

– greater that the D&C algorithm, but the both values are almost comparable,

while its computational efforts are many times less than the exact D&C algorithm's ones and almost equal (in general) with the H2 heuristic (that is better than H1 by its number of misclassifications).

5. Conclusions

We propose a fast heuristic algorithm FCS for solving the classical two-group classification problem. The main idea of the algorithm is to create consecutively several sets of "promising" observations. Then a small part of the possible combinations of observations in each set are evaluated in search for a better objective function value. The carried out simulated study (with multivariate normal distribution data sets) demonstrates that the obtained FCS solutions are very close (in respect to the number of misclassifications) to the optimal solutions while their computational efforts (pivots, pricing) are many times less.

Acknowledgment

The second author would like to thank Professor Paul Rubin, who has run his two heuristics (H1 and H2) on our simulated data sets and sent the results (included in Table 1) and Dr. Antonio Pedro Silva that sent his software for running the "Divide and Conquer" algorithm.

References

1. Abad, P., W. 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 removing the smallest number of relation from linear systems to achieve feasibility. Unpublished manuscript, 1994.
3. Asparoukhov, O., S. Danchev. Discrimination and classification in the presence of binary variables. – Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, **17**, 1997, No 1/2, 25-39.
4. Asparoukhov, O. , A. Stam. Mathematical programming formulations for two group classification with binary variables. – Annals of Operations Research, **74**, 89-112, 1997.
5. Bajgier, S., A. Hill. An experimental comparison of statistical and linear programming approaches to the discriminant problems. Decision Sciences, **13**, 1982, 604-618.
6. 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 .
7. Joachimsthaler, E., A. Stam. For approaches to the classification problem in discriminant analysis: An experimental study. – Decision Sciences, **19**, 1988, 322-333.
8. Joachimsthaler, E., A. Stam. Mathematical Programming approaches for the classification problem in two-group discriminant analysis. – Multivariate Behavioral Research, **25**, 1990, 427-454.
9. Koehler, G., S. Erenguc. Minimizing misclassifications in linear discriminant analysis. – Decision Sciences, **21**, 1990, 63-85.
10. Lam, K., Choo, E., W. Wedley. Linear goal programming in estimation of classification probability. – European Journal of Operational Research, **67**, 1993, 101-110 .
11. McLachlan, G. Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley, 1992.
12. Nath, R., W. Jackson, T. Jones. A comparison of the classical and the linear programming approaches to the discriminant analysis. – Decision Sciences, **19**, 1992, 554-563.
13. Rubin, P. A. Heuristic solution procedures for a mixed-integer programming discriminant model. – Managerial and Decision Economics, **11**, 1990, 255-266.
14. Rubin, P. Solving mixed integer classification problems by decomposition. – Annals of Operations Research, **74**, 1997, 51-64.
15. Soltysik, R., P.Yarnold. The Warmack-Gonzales algorithm for linear two-category multivariable optimal discriminant analysis. – Computers Ops. Res., **21**, 1994, No 7, 735-745.
16. Stam, A., E. Joachimsthaler. Solving the classification problem in discriminant analysis via Linear and Nonlinear Programming methods. – Decision Sciences, **20**, 1989, 285-293.
17. Stam, A., E. 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.
18. Stam, A., D. 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.
19. Stam, A. MP approaches to classification: Issues and Trends. – Annals of Operations Research, **74**, 1997, 1-36.
20. Warmack, R., R. Gonzalez. An algorithm for the optimal solution of linear inequalities and its application to pattern recognition. – In: IEEE Trans. on Computers, **C22**, 1973, 1065-1075.

Эвристическая процедура для задачи классификации в двух группах

Васил Г. Гуляшки, Огнян К. Аспарухов***

**Институт информационных технологий, 1113 София*

***Центр биомедицинского инженерства, 1113 София*

(Резюме)

Основные формулировки смесено-целочисленного программирования атакуют прямо цель – минимизирование числа ошибочных классификаций, при чём они

требуют большие вычислительные усилия. Поэтому целесообразно рассмотривать применение эвристических процедур, которые имеют сравнимую точность, но работают гораздо быстрее. В этой статии представлена такая эвристика, которая основывается на быстром комбинаторном поиске, имея в виду специфику классификационной задачи. Осуществлено симулированое исследование (с мультивариационными множествами данных, имеющих нормальное распределение). Оно показывает, что предложенная эвристика даёт решения очень близкие (по отношению числа ошибочных классификаций) к оптимальным решениям, тогда как вычислительные усилия (пивоты, прайсинг) на много раз меньше.