Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010), 30-37

2155-0182/10/$3.00

Two-Group MultiODA: A Mixed-Integer Linear Programming Solution with Bounded *M*

Robert C. Soltysik, M.S., and Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Prior mixed-integer linear programming procedures for obtaining two-group multivariable optimal discriminant analysis (MultiODA) models require estimation of the value of a parameter, *M*. A new formulation is presented which establishes a lower bound for *M*, which executes more quickly than prior formulations. A sufficient condition for the nonexistence of classification gaps and ambiguous solutions, optimal weighted classification, use of nonlinear terms, selecting an optimal subset of attributes, and aggregation of duplicate observations are discussed. When the design involves six or fewer binary attributes, MultiODA models may easily be obtained for massive samples.

.

Classification models derived via multivariable optimal discriminant analysis (MultiODA) are linear discriminant classifiers which explicitly maximize classification accuracy for a given sample.[1] Mixed-integer linear programming formulations for two-group MultiODA models require estimation of the value of a parameter, *M*, commonly defined as "a prohibitively large number."[2] If the estimated value of *M* is too low then suboptimal solutions may occur, and excessively large values of *M* will decrease computational efficiency and may introduce numerical (round-off) error.[3] We present a goal programming formulation which establishes a lower bound for *M*, and then we discuss a sufficient condition for the nonexistence of classification gaps and ambiguous solutions, weighted classification, the use of nonlinear terms, selection of optimal subsets of attributes, and aggregation of duplicate observations.

MIP45 Methodology

In a two-group linear MultiODA problem with *p* attributes and *m* observations, a set of *m* row vectors $\boldsymbol{a_i}$ is given, the components of which are $p = n$-1 observed values and a dummy value of unity. Each observation *i* is a member of either class 0 or class 1. A weight vector $\boldsymbol{x}$ is determined so that *i* is predicted to belong to class 0 when $\boldsymbol{a_i x} < 0$, or to class 1 when $\boldsymbol{a_i x} > 0$. Observation *i* is considered to be correctly classified if its predicted class membership is the same as its actual class membership, and misclassified otherwise. Solutions of interest yield maximum classification accuracy, that is,

2155-0182/10/$3.00

minimize the number of misclassified observations. This is achieved by determining $\boldsymbol{x}^*$ which satisfy the maximum number of inequalities in the system:

$\boldsymbol{a}_i\boldsymbol{x} < 0$ for observations in class 0,

$\boldsymbol{a}_i\boldsymbol{x} > 0$ for observations in class 1. (1)

This problem may be formulated as a mixed-integer linear programming model. To accomplish this, the strict inequalities in (1) are replaced with $\boldsymbol{a}_i\boldsymbol{x} \leq -\varepsilon$ or $\boldsymbol{a}_i\boldsymbol{x} \geq \varepsilon$, where $\varepsilon \geq 0$. This is necessary due to the inability of simplex-based algorithms for mixed-integer programming to handle strict inequalities (mixed-integer techniques based upon interior-point algorithms[4] may not suffer this limitation). Letting ε be strictly positive removes the ambiguity in the classification status of observations i for which $\boldsymbol{a}_i\boldsymbol{x} = 0$, but also introduces the possibility of a classification gap. It will be shown that there are conditions under which ambiguities can be removed for $\varepsilon = 0$. Consider the following model:

$$\text{MIP45: } z = \min \sum_{i=1}^{m} d_i \quad (2)$$

subject to

$$\sum_{j=1}^{n} a_{ij} (x_j^+ - x_j^-) - M_i d_i \leq -\varepsilon, \; i \in I_0 \quad (3)$$

$$\sum_{j=1}^{n} a_{ij} (x_j^+ - x_j^-) + M_i d_i \geq \varepsilon, \; i \in I_1 \quad (4)$$

$$\sum_{j=1}^{n} (x_j^+ + x_j^-) = 1 \quad (5)$$

$$x_j^+ - g_j \leq 0, \; j=1,..., n \quad (6)$$

$$x_j^- + g_j \leq 1, \; j=1,..., n \quad (7)$$

$$x_j^+, x_j^- \geq 0, \; j=1,..., n \quad (8)$$

$$g_j \in \{0,1\}, \; j=1,..., n \quad (9)$$

$$d_i \in \{0,1\}, \; i=1,...,m \quad (10)$$

where

a_{ij} is the jth component of observation $\boldsymbol{a}_i$

I_0 is the set of observations belonging to class 0

I_1 is the set of observations belonging to class 1

$$M_i = \max_j |a_{ij}| + \varepsilon \quad (11)$$

z is the number of misclassified observations.

The weight vector $\boldsymbol{x}$ is obtained by

$$x_j = x_j^+ - x_j^-, \; j=1,..., n. \quad (12)$$

Since constraints (6) and (7) ensure that not more than one of the x_j^+ and x_j^- are positive for any j, we can think of these values as the "positive" and "negative" parts of x_j, respectively. Note that $g_j = 1$ when $x_j > 0$ and $g_j = 0$ when $x_j < 0$. Also note that the g_j, along with (6), (7), and (9), may be dropped when $\varepsilon > 0$.

Constraint (5) normalizes $\boldsymbol{x}$ so that

$$\sum_{j=1}^{n} |x_j| = 1 ; \quad (13)$$

that is, the sum of the absolute values of the discriminant weights is constrained to equal one. This normalization prevents the trivial solution $\boldsymbol{x} = \boldsymbol{0}$ (when $\varepsilon > 0$), and allows us to establish a lower bound for the M_i. It is necessary for the M_i to be large enough to force compliance of the constraints (3) and (4). This is accomplished by (11). To see this, consider constraint (4). Since $\sum_j |x_j| = 1$, it is clear that

2155-0182/10/$3.00

$$\boldsymbol{a}_i\boldsymbol{x} \geq -\max_j |a_{ij}| \quad (14)$$

and

$$\boldsymbol{a}_i\boldsymbol{x} + \max_j |a_{ij}| + \varepsilon \geq \varepsilon. \quad (15)$$

Therefore, when $d_i = 1$,

$$\boldsymbol{a}_i\boldsymbol{x} + M_i d_i \geq \varepsilon. \quad (16)$$

Because the normalization (5) requires that all optimal weight vectors $\boldsymbol{x}^*$ lie on a 45° properly rotated hypercube centered at the origin, this formulation is referred to as MIP45. It may be the case that more than one solution for $\boldsymbol{d}$ may be optimal for a problem. This corresponds to the existence of multiple optimal dichotomies of predicted class membership. It is also generally true that a solution space for $\boldsymbol{x}$ of positive volume exists for each dichotomy. The issue of selecting among optimal $\boldsymbol{x}^*$ may be addressed by a number of methods, such as linear programming[5] and *a priori* decision heuristics.[6]

Resolving Classification Gaps and Ambiguities

In the above formulation, at least n - 1 of the $\boldsymbol{a}_i\boldsymbol{x}^*$ are at zero when $\varepsilon = 0$ is specified**.** From (1), it is seen that the criterion of strict separation of the classes should be met. An optimal value $z^* > 0$ in the solution of the following linear program guarantees that this separation is maintained.

LP: max $z = y$

subject to

$$\sum_{j=1}^{n} a_{ij}(b_j^+ - b_j^-) + \underline{y} \leq 0,\ \underline{i} \in \underline{I_0} \text{ and } \boldsymbol{a}_i\boldsymbol{x}^* \leq 0 \quad (17)$$

$$\sum_{j=1}^{n} a_{ij}(b_j^+ - b_j^-) - \underline{y} \geq 0,\ \underline{i} \in \underline{I_1} \text{ and } \boldsymbol{a}_i\boldsymbol{x}^* \geq 0 \quad (18)$$

$$\sum_{j=1}^{n} (b_j^+ - b_j^-) = 1 \quad (19)$$

$$b_j^+, b_j^-, y \geq 0 \quad (20)$$

$$b_j = b_j^+ + b_j^- . \quad (21)$$

This LP may be executed for each optimal dichotomy. If $z^* > 0$ is obtained, $\boldsymbol{b}^*$ is a new discriminant vector which optimizes criterion (1). Otherwise, ambiguity remains in the classification status of observations for which $\boldsymbol{a}_i\boldsymbol{b}^* = 0$: such observations should not be classified.

The advantage of establishing a lower bound for M is illustrated with an example involving discriminating between excellent versus less than excellent medical residents using information obtained during their application for residency training. Since rating applicants for residency training is a difficult, time-intensive decision-making task, a linear discriminant classifier that successfully predicts resident performance might be of great interest and utility to admissions committees.

The sample was $m = 49$ residents enrolled in a three-year internal medicine residency program.[7] The clinical performance (class) variable was based on the mean rating on an explicit 10-point scale made by residents' supervisors: a mean rating of nine or greater on this scale reflected "excellent" (or better) clinical performance (class = 1, $m_1 = 27$), and a mean rating of less than nine reflected less than excellent clinical performance (class = 0; $m_0 = 22$). The n - 1 = 3 application information variables (attributes) included medical board scores, faculty evaluations (a composite measure reflecting ratings of letters of recommendation and medical school grading system),

2155-0182/10/$3.00

and academic distinction (a composite measure reflecting honors attained in medical school and medical school status).

The computer resources required to solve this problem using MIP45 versus Stam and Joachimsthaler[8] was compared (other prior formulations were slower). For MIP45, ε was set at 0. For Stam and Joachimsthaler, values of 1, 10, 100, and 1000 were used for M, and a value of 1 for ε.[9] All formulations were solved on an IBM 3090/300 computer running SAS/OR.[10] As seen in Table 1, except when M = 1, Stam and Joachimsthaler required more computational effort (CPU time, pivots, and integer branches) than did MIP45. Using M = 1, ε = 1 in Stam and Joachimsthaler resulted in a useless solution, and using M = 10 or 100 resulted in suboptimal solutions of (3). Since a decision-maker using M = 10 or M = 100 would have no direct evidence that these solutions were suboptimal, it would also be unclear whether the solution attained by Stam and Joachimsthaler (or other unbounded formulations) using M = 1000 was optimal. In contrast, since the value of z* attained in LP was positive, a decision-maker using MIP45 to solve this problem would be certain that the solution was unambiguously optimal: a clear advantage.

TABLE 1

Illustration of Computational Resources Needed by MIP45 Versus Stam and Joachimsthaler[8] to Solve a Problem with 49 Observations and Three Attributes, Using SAS/OR run on an IBM 3090/300 Computer

Formulation	M	ε	Objective Value	CPU Seconds	Integer Branches	Pivots
Stam	1	1	29	1.1	0	31
Stam	10	1	17	131.8	8,629	36,607
Stam	100	1	15	276.7	19,755	89,564
Stam	1000	1	14	268.4	14,549	57,351
MIP45	LB	0	14	48.0	2,896	15,333

Note: For MIP45 the M_i were set at their lower bounds (LB). For solutions resulting in the optimal value of 14 misclassifications, model coefficients for board scores and faculty evaluation were positive, and the coefficient for academic distinction was negative. For MIP45, z* = .00439.

Weighted Classification

Rather than weighting each observation equally, we consider weighting each case in (2) by a positive scalar c_i. This is significant for two reasons. First, the c_i may represent the cost of misclassifying observation i. In this case an optimal solution would minimize the cost of misclassification (or, equivalently, maximize the return of correct classification) for the sample. Second, the c_i may represent factors which balance the number of class 0 and class 1 observations when these are not equal. In this case an optimal solution would maximize the number of

2155-0182/10/$3.00

correct classifications weighted by population membership in each class. An example would be $c_i = 1/m_0$ for observations in class 0, and $c_i = 1/m_1$ for observations in class 1, where m_0 and m_1 are the number of observations in categories 0 and 1, respectively. This latter weighting scheme is particularly useful in badly imbalanced applications for which $m_0 >> m_1$, or visa versa: use of such "priors weights" forces the model to classify observations from both classes accurately, and inhibits the identification of degenerate models which classify all observations into a single class category.

Adding Nonlinear Terms as Attributes

Here we generalize the notion of maximum pattern classification accuracy achieved by separating hyperplanes to sets of nonlinear separating surfaces. For example, consider quadratic surfaces in p-measurement space of the form:

$$\sum_j a_{ij} \underline{x}_j + \sum_{k \le p} \sum_{l \le k} a_{ik} a_{il} x_{kl} + a_{in} x_n \qquad (22)$$

for all i. The MultiODA solution can be attained by augmenting the a_j and x in the MIP45 model by the interaction terms in (22). This solution produces a weight vector $\boldsymbol{x}$ which yields the minimum number of misclassifications achievable by a quadratic separating surface. This process may be applied to any nonlinear discriminant function which is linear in the parameters of the measurement space.

Optimal Attribute Subset Selection

In the foregoing we have assumed that all p attributes are included in the MultiODA model. However, we may wish to select a subset of $k < p$ attributes for the application of the model. For example, imagine an application involving 50 observations and ten attributes. In order to identify a model that may generalize if used to classify independent random samples, we may wish to maintain a minimum observation-to-attribute ratio of 10-to-1, so a maximum of five of the ten potential attributes may be used. Of all possible 5-attribute models, which yields maximum accuracy? Optimal attribute subset selection methodology can be incorporated in the MIP45 model by defining n zero-one variables q_j and including the following constraints:

$$\underline{x}_j^- - q_j \le 0, \; j=1,..., n, \qquad (23)$$

$$g_j + q_j \le 1, \; j=1,..., n, \qquad (24)$$

and

$$\sum_{j=1}^{n} g_j + \sum_{j=1}^{n} q_j = k. \qquad (25)$$

In an optimal solution to such a MultiODA model, measurement $\underline{j}$ is selected for inclusion only if $g_j + q_j = 1$. The number of misclassifications obtained is the fewest achievable in any k-dimensional subspace of the original p-dimensional measurement space.

Aggregation of Duplicate Observations

If duplicate observations occur in the data set (i.e., two or more observations have the same value for every attribute measurement), the following procedure may be used to aggregate the duplicate observations into a single observation, reducing the size of the overall problem. The resulting problem is equivalent to the original one, with m' observations, and objective value $z + v$.

1. $m' := m : s_0 = 0: s_1 = 0: v := 0$
2. **for each** $i = 1, \ldots, m'$
3. **for each** $j < i$
4. **if** $a_i = a_j$ **then**
5. **if** $i \in \underline{I}_0$ **then** $s_0 := s_0 + c_i$ **else** $s_1 := s_1 + c_i$
6. remove observation i from list : $m' := m' - 1$
7. **end if**

2155-0182/10/$3.00

8. **next** j, i
9. **for each** $i = 1, \ldots, m'$
10. **if** $s_0 > s_1$ **then**
11. $w_j := s_0 - s_1 : v := v + s_1$
12. **else if** $s_1 > s_0$ **then**
13. $w_j := s_1 - s_0 : v := v + s_0$
14. **else**
15. $v := v + s_0$: remove observation i from list : $m' := m' - 1$
16. **end if**
17. **next** i

This procedure is particularly useful when $\boldsymbol{a}_j$ is a zero-one vector (all attributes are binary). Here all the patterns lie on the vertices of the p-dimensional unit hypercube. If more than one pattern lies on some vertex, then by using the above procedure we may obtain a weighted MIP45 model equivalent to the original model, but with fewer constraints. If the number of original patterns m is large relative to the number of attributes p, a significant reduction in the size of the model may be obtained. For instance, regardless of the value of m, if p=8 then we end up with no more than $2^8 = 256$ constraints of type (6) in the model. Since the number of constraints is independent of m, extremely large problems may be solved with this procedure, provided p is moderately small.

In order to illustrate the potential solution efficiency gained by using this special purpose algorithm for problems involving entirely binary data, we ran 30 Monte Carlo experiments. In each experiment there were five binary attributes, such that the total possible number of different profiles was $2^5 = 32$. Values on each attribute were determined separately for each observation on the basis of a random uniform number between 0 and 1: numbers < 0.5 were assigned the value of 0, and numbers $\geq$ 0.5 were assigned the value of 1. We ran five balanced ($m_0 = m_1$) experiments for each total sample size of 50, 100, 1000, 10^4, 10^5, and 10^6 total observations. All formulations were solved on an IBM 3090/300 computer running SAS/OR. As seen in Table 2, as the number of observations increased: (a) the number of distinct profiles increased toward its theoretical upper bound (the theoretical upper bound was achieved in all of the problems involving 10^6 observations, and in four of the five problems involving 10^5 observations); (b) the misclassification rate increased towards its theoretical upper bound (i.e., for a balanced design with an even number of observations, the theoretical upper bound for the number of misclassifications is one less than one-half of the total number of observations); and (c) the mean number of CPU seconds required to solve the problem was approximately twenty seconds for problems with 1000 or *more* total observations.

TABLE 2

Results of Monte Carlo Experiments for Binary Data: Five Random Attributes

Number of Observations	Number of Profiles	Number (%) of Misclassifications	CPU Seconds
50	19	14 (28%)	4.5
50	23	13 (26%)	5.5
50	20	16 (32%)	8.4

2155-0182/10/$3.00

50	23	14 (28%)	12.5
50	21	12 (24%)	1.7
100	30	27 (27%)	14.2
100	26	34 (34%)	9.0
100	25	43 (43%)	10.5
100	24	37 (37%)	7.5
100	20	33 (33%)	3.0
1000	32	432 (43%)	17.8
1000	30	445 (44%)	25.1
1000	31	449 (45%)	16.4
1000	31	460 (46%)	24.3
1000	31	454 (45%)	19.2
10000	29	4870 (49%)	12.3
10000	31	4838 (48%)	23.8
10000	31	4842 (48%)	24.9
10000	29	4828 (48%)	11.9
10000	31	4839 (48%)	9.2
100000	32	49545 (50%)	14.5
100000	32	49532 (50%)	21.6
100000	31	49526 (50%)	6.3
100000	32	49475 (49%)	25.2
100000	32	49376 (49%)	16.8
1000000	32	498331 (50%)	24.3
1000000	32	498759 (50%)	17.2
1000000	32	498450 (50%)	32.5
1000000	32	497861 (50%)	4.5
1000000	32	498837 (50%)	16.8

Discussion

MIP45 solves two problems common to prior goal programming formulations of two-group MultiODA: *M* is automatically set at its lower bound, and it is possible to determine whether classification gaps or ambiguities exist. Collateral benefits of MIP45 include its greater computational efficiency and solution speed relative to prior formulations, particularly for applications involving binary attributes.

This study contrasted the computational characteristics of the MIP45 formulation of the MultiODA problem to the formulation of Joachimsthaler and Stam (see Table 1). Other mixed-integer programming formulations have appeared more recently. Rubin developed a decomposition technique to solve the Multi-ODA problem.[11] Silva and Stam developed a partitioning method for MultiODA which was reported to compare favorably with MIP45.[12] Pfetsch developed a technique to optimize

2155-0182/10/$3.00

irreducible inconsistent subsystems (IIS) of linear inequalities in order to determine a maximum feasible subsystem of these inequalities.[13] Finally, Bremner and Chen developed a MIP formulation for the halfspace depth problem which uses IIS cuts in a branch-and-cut algorithm.[14] We eagerly anticipate computational comparisons between these formulations.

References

[1]Yarnold PR, Soltysik RC, Martin GJ. Heart rate variability and susceptibility for sudden cardiac death: an example of multivariable optimal discriminant analysis. *Statistics in Medicine* 1994, 13:1015-1021.

[2]Joachimsthaler EA, Stam A. Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioral Research* 1990, 25:427-454.

[3]Gehrlein WV. General mathematical programming formulations for the statistical classification problem. *Operations Research Letters* 1986, 5:299-304.

[4]Karmarkar N. A new polynomial time algorithm for linear programming. *Combinatorica* 1984, 4:373-395.

[5]Koehler GJ, Erenguc SS. Minimizing misclassifications in linear discriminant analysis. *Decision Sciences* 1990, 21:63-74.

[6]Yarnold PR, Soltysik RC. Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences* 1991, 22:739-752.

[7]Curry RH, Yarnold PR, Bryant FB, Martin GJ, Hughes RL. A path analysis of medical school and residency performance: implications for housestaff selection. *Evaluation in the Health Professions* 1988, 11:113-129.

[8]Stam A, Joachimsthaler EA. A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research* 1990, 46:113-122.

[9]Bajgier SM, Hill AV. An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences* 1982, 13:604-612.

[10]Cornell R, Luginbuhl RC, Yeo C. *SAS/OR user's guide, version 6.* SAS Institute, Durham, NC, 1989.

[11]Rubin PA. Solving mixed-integer classification problems by decomposition. *Annals of Operations Research* 1997, 74:51-64.

[12]Silva APD, Stam A. A mixed-integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. *Annals of Operations Research* 1997, 74:129-157.

[13]Pfetsch ME. Branch-and-cut for the maximum feasible subsystem problem. *SIAM Journal on Optimization* 2008, 19:21-38.

[14]Bremner D, Chen D. A branch and cut algorithm for the halfspace depth problem. 2009: arXiv:0910.1923v1.

Author Notes

This research was supported in part by a grant from the NSF (SES-8822337), and an early version of the paper was invited for presentation at the TIMS/ORSA Joint National Meeting, Chicago, 1993. Appreciation is extended to the computer center of the University of Illinois at Chicago, for computer resources used in this research. Mail correspondence to authors at: Optimal Data Analysis, 1220 Rosecrans Street, Suite 330, San Diego, CA 92106. Send Email to: Journal@OptimalDataAnalysis.com.