Solving a Mixed-Integer Programming Formulation of a Multi-Category Constrained Discrimination Model

J Paul Brooks[1] and Eva K Lee[2]
[1] Virginia Commonwealth University, [2]Georgia Institute of Technology

1 Abstract

In classification, even if a Bayes-optimal rule has been developed, intergroup misclassification rates may be higher than desirable. We consider a two-stage model for multi-category constrained discrimination in which limits on misclassification rates of training observations may be pre-specified. The mechanism by which the misclassification limits are satisfied is a rejection option, also known as a reserved judgment group, for observations not demonstrating properties of membership to any of the groups.

The first stage of the constrained discrimination model involves estimating conditional group density function values for training observations, and the second stage requires the solution of a mixed-integer program (MIP). The MIP is used to estimate the parameters that characterize an optimal classification rule.

The MIP is difficult to solve due to the formulation of constraints wherein certain variables are equal to the maximum of linear functions. Solution methods for the MIP are presented, including techniques for generating and exploiting edges of the conflict graph. Improvement in computation times over industry-standard software is demonstrated.

2 Introduction

Classification is a fundamental task in machine learning where rules are developed for the allocation of observations, represented by vectors of measurable attributes, to one of a number of discrete groups. Commonly employed methods for classification include classification and regression trees [5], support vector machines, neural networks, and linear discriminant functions. Typically, a set of training observations whose group memberships are known are used to develop rules for the classification of unknown group membership in a process called supervised learning.

The Bayes optimal rule for a given data set is the rule that maximizes the probability of correct classification of observations of unknown group membership; or equivalently, minimizes the probability of misclassification (see [10]). For an observation x, the Bayes optimal rule allocates an observation to the group h such that $h = \arg\max_g \{\pi_g f_g(x), g \in \mathcal{G}\}$ where π_g is the prior probability that an observation belongs to group g, $f_g(x) = f(x|g)$ is the conditional density function for the event that an observation exhibits values x given that it belongs to group g, and $\mathcal{G}$ is the set of groups. The product of the prior probability and conditional density function for an observation-group pair is the posterior probability that the observation belongs to the group. Therefore, the Bayes optimal rule allocates observations to groups for which the posterior probability is the maximum.

The Bayes optimal rule is rarely obtainable, and classification methods are often evaluated by aspects of their convergence to the Bayes optimal rule [10, 11]. Even when the Bayes optimal rule is known, and the total probability of correct classification is maximized, the rate of misclassification between pairs of groups is often higher than desired. Anderson [1] notes this phenomenon, and presents a modification of the Bayes optimal rule that will provably maximize the probability of correct classification while satisfying limits on the probabilities of misclassification of observations between pairs of groups. An *Anderson optimal rule* is defined by linear modifications to the posterior probabilities called *modified posterior probabilities*. The mechanism by which the constraints are satisfied is through the introduction of a reserved judgment, or rejection, group. The reserved judgment group is an artificial group where observations can be placed if the modified posterior probabilities are low for all groups.

A classification rule with a rejection option is a *constrained discrimination rule*. Forney [15] and Chow [8] discuss the form of rules that minimize the rejection rate given an error rate. Györfi et al. [18] suggest a stochastic approximation procedure for finding an optimal classification rule for this objective and constraint. The approach in this work will be to maximize correct classification while constraining error rates, which is equivalent to minimizing the rejection rate with misclassification limits. Other approaches to constrained discrimination have been investigated as well [29, 19, 6].

Gallagher et al. [16] present three computational models for estimating the parameters of an Anderson optimal rule. These models involve the solution of computationally intensive mixed-integer programs. The difficulty in solving the MIPs led to the development of a linear programming approximation [25]. The linear programming approximation has been used successfully in prediction of ultrasound-mediated disruption of cell membranes [24], automated planning volume definition in soft-tissue sarcoma adjuvant brachytherapy [23], and genomic pattern recognition in human cancer [13, 14].

This paper describes some of the solution methods for the mixed-integer programming formulation given in part in [16]. We will operate under the assumption that the Bayes optimal rule is known, and discuss the estimation of parameters of the Anderson optimal rule. In particular, methods for determining edges of the conflict graph of the mixed-integer program are given. The conflict graph corresponds to a node packing relaxation of a mixed-integer program. Conflict graphs and their use in solving combinatorial optimization problems are described in [28, 26, 20, 17, 21, 4, 3, 2]. The conflict graph can be exploited to derive cuts in a branch and bound framework.

3 Formulation

Input to the constrained discrimination model proposed in [16] consists of the following: (1) A set of groups/categories/classes $\mathcal{G}$, (2) estimates of prior probabilities π_h describing the likelihood that an observation with unknown attribute values belongs to group $h \in \mathcal{G}$, (3) a set of observations/entities/data points $\mathcal{N}_h$ for each group $h \in \mathcal{G}$ such that the observations in $\mathcal{N}_h$ are known to belong to group h, (4) the conditional group density functions evaluated at the points representing each of the observations (the conditional group density function for observation j with known group membership g for allocation to group h is written as $f_h(x^{gj})$), and (5) misclassification rate limits $\alpha_{hg} \in [0, 1]$ that represent the highest proportion of observations from group h that can be allocated to group g.

Anderson [1] shows that an optimal rule for maximizing correct classification subject to pre-specified limits on misclassification allocates observations to groups based on modified posterior probabilities. Let L_{hgj} be the modified posterior probability that observation j with group membership g is allocated to group h. The optimal rule is determined by nonnegative $|\mathcal{G}|(|\mathcal{G}| - 1)$ λ_{ih} coefficients that relate to the modified posterior probabilities by the following formula

$$L_{hgj} = \pi_h f_g(x^{gj}) - \sum_{i \in \mathcal{G} \setminus h} f_i(x^{gj}) \lambda_{ih}$$

Note that the modified posterior probabilities are not probabilities in that they are not required to sum to 1, or even be nonnegative. The name *modified posterior probability* results from the fact that the quantity indicates the likelihood of an event; namely, the allocation of an observation to a group. The optimal rule presented by Anderson allocates an observation to the group for which the modified posterior probability is greatest unless $L_{hgj} < 0$ for all h in which case the observation is placed in the reserved judgment region.

The estimation of the λ_{ih} values based on a sample of observations with known group membership is performed using math programming. Let $u_{hgj} = 1$if entity x^{gj} is allocated to group h and 0 otherwise, and $y_{gj} = \max\{0, L_{h'gj} : h \in \mathcal{G}\}$. An MIP formulation of the objective and constraints is then

$$\begin{array}{lrcll}
\text{maximize} & \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} u_{ggj} & & & \\
\text{subject to} & y_{gj} & \leq & M(1 - u_{0gj}) & g \in \mathcal{G},\ j \in \mathcal{N}_g \quad (1) \\
& y_{gj} & \geq & \epsilon u_{hgj} & g,\ h \in \mathcal{G},\ j \in \mathcal{N}_g \quad (2) \\
& y_{gj} - L_{hgj} & \leq & M(1 - u_{hgj}) & g,\ h \in \mathcal{G},\ j \in \mathcal{N}_g \quad (3) \\
& y_{gj} - L_{hgj} & \geq & \epsilon(1 - u_{hgj}) & g,\ h \in \mathcal{G},\ j \in \mathcal{N}_g \quad (4) \\
& \sum_{0, h \in \mathcal{G}} u_{hgj} & = & 1 & g \in \mathcal{G},\ j \in \mathcal{N}_g \quad (5) \\
& \sum_{j \in \mathcal{N}_g} u_{hgj} & \leq & \lfloor \alpha_{gh} n_g \rfloor & h,\ g \in \mathcal{G},\ h \neq g \quad (6) \\
& -\infty < L_{hgj} < \infty,\ y_{gj} \geq 0,\ \lambda_{ih} \geq 0,\ u_{hgj} \in \{0, 1\} & & &
\end{array}$$

This model is almost identical to Model 1 formulated by [16]. Note that for appropriately chosen M and ϵ, the constraints (1)-(4) allocate observations to the group for which the modified posterior probability is maximum. Constraints (5) require that each observation is allocated to exactly one group, and constraints

(6) enforce the misclassification limits by requiring that no greater proportion than α_{gh} of the observations belonging to group g are misclassified as belonging to group h. The existence of the M and ϵ in a large number of constraints is conducive to an extremely ill-conditioned coefficient matrix.

4 Computational Techniques and Tests

The focus of our computational approach for difficult MIP instances is finding and exploiting the conflict graph. The conflict graph of a mixed-integer program is a graph with one node corresponding to each binary integer variable. An edge exists between variables/nodes x_i and x_j if $x_i + x_j \leq 1$ is valid for the associated MIP. Certain structures, including maximal cliques and odd holes, in the conflict graph correspond to stronger valid inequalities than the edge constraints (see [21, 4, 3]). In fact, the edges of the conflict graph of an integer program form a node packing relaxation of the convex hull of integer feasible solutions.

For the MIP in Section 3, we can show that determining if an edge is in the conflict graph is equivalent to showing that a system of 4 constraints and $|\mathcal{G}|(|\mathcal{G}-1|)$ variables is infeasible. An edge corresponds to a constraint of the form $u_{amn} + u_{bst} \leq 1$, which is a logical condition that observation n from group m cannot be allocated to group a in the same solution that observation t from group s can be allocated to group b.

To demonstrate the computational improvement, we used 5 3-group and 4 5-group real world data sets. With the exception of one data set, all of the data sets are from the UCI machine learning database repository [9]. The remaining data set consists of cell motility data [30]. The proportion of training observations from each group served as prior probabilities, and the conditional group density function values were estimated by assuming a multivariate normal distribution of the data with a common covariance matrix. Splus 6.0 for Linux/Unix is used for calculating the conditional group density function values.

The branching strategy incorporates information from the conflict graph and misclassification constraints. For example, when a binary integer variable is set to 1, all variables that are adjacent in the conflict graph are set to 0. Noting that the λ_{ih} variable values determine all other variable values, a heuristic is used at nodes to generate integer feasible solutions at nodes in the branch and bound tree. At each fractional node, the λ_{ih} values that optimize the LP subproblem are used to calculate the modified posterior probabilities, and then the values of the integer variables. If the integer variables satisfy the misclassification constraints and the objective value is larger than the incumbent objective value, then the solution is stored as the new incumbent.

Instances of the classification model are solved using CPLEX Callable Library V8.1 (CPLEX). For comparison, the CPLEX solver is then enhanced with the specialized cuts, branching strategy, and heuristic based on our analysis. The entire conflict graph is generated and stored before optimization. The conflict graph is searched at nodes in the branch and bound tree for maximal clique and odd hole cuts. The Cliquer [27] routines are used to search for maximal cliques in the fractional conflict graph and a breadth-first search [21, 3] is used to search for odd holes.

Several combinations of settings available in CPLEX were tested. To summarize, strong branching, aggressive probing, aggressive generation of clique cuts, and strict tolerance levels are employed for the results presented here. All problems were solved on machines with 2×2.4GHz Intel Xeon processors and 2GB RAM. CPLEX was set to terminate after 7,200 CPU seconds.

Each data set was tested with misclassification limits of 0, 5, 15, and 100%. The last set of limits corresponds to having no misclassification limits. Statistically, solving the problem with no misclassification limits is of little use; however, the problem is a special case of the problem with misclassification limits and is of interest when evaluating computational performance.

A comparison of computational performance on 5-group data sets is in Table 4. The first two columns identify the problem instance by data set and misclassification rate limits. Columns 3 through 8 summarize the performance of CPLEX alone, and columns 9 through 14 summarize the performance of the Enhanced Solver. Columns 3 and 9, *Best obj. value*, contain the objective value of the best integer feasible solution from the respective solvers. Columns 4 and 10, *Gap rem (%)* contain the percentage of integrality gap remaining, which is calculated as $(z_{UB}^{LP} - z^{IP})/(z_{root}^{LP} - z^{IP}) \times 100$, where z^{IP} is the optimal objective value of the mixed-integer program. If z^{IP} is unknown, the objective value associated with the best known integer feasible solution is used. The value z_{UB}^{LP} is the best objective value among active LP subproblems upon termination. The value z_{root}^{LP} is the optimal objective value of the linear program relaxation at the root of the branch and bound tree. Columns 5 and 11, *Time to best (s)* include the time in CPU seconds to find an

integer feasible solution with objective at least as good as that obtained by running CPLEX alone. Columns 6 and 12, *Nodes to best*, *Nodes to CPLEX best*, report the number of branch-and-bound nodes solved to obtain a solution with objective value at least as good at that obtained by CPLEX alone. Columns 7, 13, 9, 14 record the termination statistics of the two solvers. *Time elapsed (s)* records the CPU seconds elapsed upon termination of each solver, either at proven optimality, or when the process reaches the maximum allowed time of 7,200 CPU seconds. *Nodes solved* reports the number of branch-and-bound nodes solved upon the termination condition. Columns 15 and 16 summarize the improvement in time and nodes, respectively, that the Enhanced Solver provides. Column 15 is calculated as the ratio of the time required by CPLEX to arrive at the best integer feasible solution to the respective time required by the Enhanced Solver. Similarly, Column 16 reports the associated ratio of the number of branch-and-bound nodes solved.

All of the 3-group problems are solved in less than 1 CPU minute by both CPLEX alone and the Enhanced Solver (data not shown). The Enhanced Solver finds feasible solutions to all 16 5-group problems, solving 9 to optimality, whereas CPLEX alone finds feasible solutions to only 9 problems, solving 7 to optimality. Of the instances not solved to optimality by both solvers, the Enhanced Solver terminates with better optimal solutions and smaller integrality gaps remaining (reduction ranges from 15-32%). Among the 9 instances in which CPLEX obtains an IP feasible solution, in all but one instance, the Enhanced Solver demonstrates an average of over 10-fold time improvement and 20-fold reduction in nodes solved compared to time and node statistics obtained by running CPLEX alone to arrive at the same IP objective value.

5 Conclusions

We have presented preliminary results that indicate that by combining cutting planes from conflict graphs, specialized branching scheme, and tailored heuristic within a branch-and-bound framework, significant improvement can be made over CPLEX in the solution of MIPs associated with a constrained discrimination model. In addition to the reduction in solution time and number of branch-and-bound nodes required to solve, and improvement in the integrality gap upon termination, a clear advantage of the Enhanced Solver is the return of feasible integer solutions to all real instances. In [7], we describe the theoretical aspects of the MIP instances and present algorithms for rapid generation of edges of the conflict graph and conflict k-hypergraphs[12, 22] to a wider class of real instances.

6 Acknowledge

This research is partially support by a grant from the National Science Foundation. The CPLEX license is provided to us by an ILOG software license grant.

References

[1] J.A. Anderson. Constrained discrimination between k populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31:123–139, 1969.

[2] Alper Atamtürk, George L. Nemhauser, and Martin W.P. Savelsbergh. Conflict graphs in solving integer programming problems. *European Journal of Operations Research*, 121:40–55, 2000.

[3] Robert E. Bixby and E. K. Lee. Solving a truck dispatching scheduling problem using branch-and-cut. *Operations Research*, 1998.

[4] R. Borndörfer. *Aspects of set packing, partitioning and covering.* PhD thesis, Technischen Universität Berlin, Berlin, Germany, 1997.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth and Brooks/Cole, 1984.

[6] J.D. Broffit, R.H. Randles, and R.V. Hogg. Distribution-free partial discriminant analysis. *Journal of the American Statistical Association*, 71:934–939, 1976.

[7] J.P. Brooks and E. K. Lee. Computing a multi-category constrained discrimination rule via mixed-integer programming and combinatorial optimization. Working paper.

[8] C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 1970.

Table 1: **CPLEX vs. Enhanced Solver.** CPLEX with cliques, GUBs, strong branching, compared to Enhanced Solver that includes maximal clique cuts, odd hole cuts, variable fixing, a user-implemented branching scheme, and a user-implemented heuristic. The optimal objective values for problems *cle-0.05* and *cle-0.15* have not been verified, herein the best-known objective values are used in calculating the percentage of integrality gap remaining for these instances. The data sets are 5-group problems.

Problem		CPLEX						Enhanced Solver						Improvement ($\frac{CPLEX}{Enhanced}$)	
				Best IP Obj.		Termination				Best CPLEX IP Obj.		Termination			
Data	α_{ij}	Best obj. value	Gap rem. (%)	Time elapsed (s)	Nodes solved	Time elapsed (s)	Nodes solved	Best obj. value	Gap rem. (%)	Time to CPLEX best (s)	Nodes to CPLEX best	Time elapsed (s)	Nodes solved	Time to CPLEX best	Nodes to CPLEX best
va	0.00	24	0.0	8	230	7	246	24	0.0	1	24	1	50	8.0	9.6
va	0.05	27	0.0	12	288	12	288	27	0.0	1	6	1	35	12.0	48.0
va	0.15	N/A	21.2	N/A	N/A	7200*	37962	49	0.0	N/A	N/A	847	7030	N/A	N/A
va	1.00	52	0.0	569	6092	757	10513	52	0.0	20	178	111	2257	28.5	34.2
swi	0.00	14	0.0	15	385	20	587	14	0.0	2	21	9	393	7.5	18.3
swi	0.05	30	0.0	2622	31223	6988	92065	30	0.0	187	1084	1819	14792	14.0	28.8
swi	0.15	N/A	44.0	N/A	N/A	7200*	14202	46	29.4	N/A	N/A	7200*	19344	N/A	N/A
swi	1.00	55	34.2	6044	53546	7200*	68330	55	0.0	995	9616	5640	53286	6.0	5.6
hun	0.00	33	0.0	25	249	49	633	33	0.0	3	2	9	96	8.3	124.5
hun	0.05	N/A	64.6	N/A	N/A	7200*	10786	104	43.2	N/A	N/A	7200*	8202	N/A	N/A
hun	0.15	N/A	66.1	N/A	N/A	7200*	5594	104	34.4	N/A	N/A	7200*	5537	N/A	N/A
hun	1.00	176	36.4	266	340	7200*	8216	177	20.9	1142	1302	7200*	23992	0.2	0.3
cle	0.00	70	0.0	185	1063	200	1208	70	0.0	21	146	107	1170	8.8	7.3
cle	0.05	N/A	56.2	N/A	N/A	7200*	9263	84	50.0	N/A	N/A	7200*	5892	N/A	N/A
cle	0.15	N/A	79.2	N/A	N/A	7200*	2747	84	51.2	N/A	N/A	7200*	1738	N/A	N/A
cle	1.00	N/A	47.8	N/A	N/A	7200*	5682	185	25.0	N/A	N/A	7200*	13601	N/A	N/A

*Not solved to optimality (terminated after 7200 CPU seconds).

[9] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64:304–310, 1989.

[10] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2001.

[12] Todd Easton, Kevin Hooker, and E.K. Lee. Facets of the independent set plytope. *Mathematical Programming, Series B*, 98:177–199, 2003.

[13] F.A. Feltus, E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. Predicting aberrant CpG island methylation. *Proceedings of the National Academy of Sciences*, 100:12253–12258, 2003.

[14] F.A. Feltus, E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. Dna motifs associated with aberrant CpG island methylation. *Genomics*, 87:572–579, 2006.

[15] Jr. G. David Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14:206–220, 1968.

[16] Richard J. Gallagher, E. K. Lee, and David A. Patterson. Constrained discriminant analysis via 0/1 mixed integer programming. *Annals of Operations Research*, 74:65–88, 1997.

[17] M.C. Golumbic, D. Rotem, and J. Urritia. Comparability graphs and intersection graphs. *Discrete Mathematics*, 43:37–46, 1983.

[18] L. Györfi, Z. Györfi, and I. Vajda. Bayesian decision with rejection. *Problems of Control and Information Theory*, 8:445–452, 1979.

[19] J.D.F. Habbema, J. Hermans, and A.T. Van Der Burgt. Cases of doubt in allocation problems. *Biometrika*, 61:313–324, 1974.

[20] E.L. Johnson and M.W. Padberg. Degree-two inequalities, clique facets, and biperfect graphs. *Annals of Discrete Mathematics*, 16:169–187, 1982.

[21] E. K. Lee. *Solving a truck dispatching scheduling problem using branch-and-cut*. PhD thesis, Computational and Applied Mathematics, Rice University, Houston, Texas, 1993.

[22] E. K. Lee and S. Maheshwary. Conflict hypergraphs in integer programming. Technical report, Georgia Institute of Technology, 2004.

[23] E. K. Lee, Albert Y.C. Fung, J. Paul Brooks, and Marco Zaider. Automated planning volume definition in soft-tissue sarcoma adjuvant brachytherapy. *Biology in Physics and Medicine*, 47:1891–1910, 2002.

[24] E. K. Lee, Richard J. Gallagher, Ann M. Campbell, and Mark R. Prausnitz. Prediction of ultrasound-mediated disruption of cell membranes using machine learning techniques and statistial analysis of acoustic spectra. *IEEE Transactions on Biomedical Engineering*, 51:1–9, 2004.

[25] E. K. Lee, Richard J. Gallagher, and David A. Patterson. A linear programming approach to discriminant analysis with a reserved-judgment region. *INFORMS Journal on Computing*, 15:23–41, 2003.

[26] George Nemhauser and L. Trotter. Properties of vertex packing and independence system polyhedra. *Mathematical Programming*, 6:48–61, 1974.

[27] Sampo Niskanen and Patrick R.J. Östergård. Cliquer user's guide, version 1.0. Technical Report Tech. Rep. T48, Communications Laboratory, Helsinki University of Technology, Espoo, Finland, 2003.

[28] M. Padberg. On the facial structure of set packing polyhedra. *Mathematical Programming*, 5:199–215, 1973.

[29] C.P. Quesenberry and M.P. Gessaman. Nonparametric discrimination using tolerance regions. *Annals of Mathematical Statistics*, 39:664–673, 1968.

[30] Adele Hart Wright. *The role of integrins in the differential upregulation of tumor cell motility by endothelial extracellular matrix proteins*. PhD thesis, Georgia Institute of Technology, December 1999.