Hybrid Misclassification Minimization

Chunhui Chen* & O. L. Mangasarian*

Mathematical Programming Technical Report 95-05
February 1995–Revised July & August 1995

Abstract

Given two finite point sets $\mathcal{A}$ and $\mathcal{B}$ in the n-dimensional real space R^n, we consider the NP-complete problem of minimizing the number of misclassified points by a plane attempting to divide R^n into two halfspaces such that each open halfspace contains points mostly of $\mathcal{A}$ or $\mathcal{B}$. This problem is equivalent to determining a plane $\{x \mid x^T w = \gamma\}$ that maximizes the number of points $x \in \mathcal{A}$ satisfying $x^T w > \gamma$, plus the number of points $x \in \mathcal{B}$ satisfying $x^T w < \gamma$. A simple but fast algorithm is proposed that alternates between (i) minimizing the number of misclassified points by translation of the separating plane, and (ii) a rotation of the plane so that it minimizes a weighted average sum of the distances of the misclassified points to the separating plane. Existence of a global solution to an underlying hybrid minimization problem is established. Computational comparison with a parametric approach to solve the NP-complete problem indicates that our approach is considerably faster and appears to generalize better as determined by tenfold cross-validation.

1 Introduction

A fundamental problem in machine learning is that of discriminating between two given point sets $\mathcal{A}$ and $\mathcal{B}$ in the n-dimensional real space R^n. This is typically achieved by constructing a plane

$$x^T w = \gamma, \tag{1}$$

such that

$$\begin{array}{ll} x^T w > \gamma & \text{for } x \in \mathcal{A} \\ x^T w < \gamma & \text{for } x \in \mathcal{B}. \end{array} \tag{2}$$

Here w is the normal to the plane and $\dfrac{|\gamma|}{\|w\|}$ is the Euclidean distance from the origin to the plane. In general it is not possible to satisfy (2) except in the special case when the convex hulls of $\mathcal{A}$ and $\mathcal{B}$ do not intersect. Thus, one resorts in the general case to minimizing some error criterion in the satisfaction of (2). The simplest such criterion is to use linear programming in order to construct a

*Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, email: *chunhui@cs.wisc.edu, olvi@cs.wisc.edu.* This material is based on research supported by Air Force Office of Scientific Research Grant F49620-94-1-0036 and National Science Foundation Grant CCR-9322479.

plane (1) that minimizes a weighted average of the sum of the distances of the misclassified points to the plane [7, 2] as follows:

$$\min_{w,\gamma,y,z} \left\{ \frac{e^T y}{m} + \frac{e^T z}{k} \mid Aw + y \geq e\gamma + e,\ Bw - z \leq e\gamma - e,\ y \geq 0,\ z \geq 0 \right\} \tag{3}$$

Here the rows of the matrices $A \in R^{m \times n}$ and $B \in R^{k \times n}$ represent the m points in $\mathcal{A}$ and the k points in $\mathcal{B}$ respectively, while e is a vector of ones of appropriate dimension. The objective function of (3) represents the sum of the average distances, multiplied by $\|w\|$, of the misclassified points in $\mathcal{A}$ to the plane $x^T w = \gamma + 1$ and of the misclassified points of $\mathcal{B}$ to the plane $x^T w = \gamma - 1$. If the convex hulls of $\mathcal{A}$ and $\mathcal{B}$ are disjoint, then there are no misclassified points and the linear program (3) yields a zero minimum. However in the general case of intersecting convex hulls, the linear program (3) obtains an approximate separating plane that minimizes an average sum of distances of misclassified points as described above. However this criterion for discrimination may not minimize the actual number of the misclassified points. The problem of constructing a plane (1) such that the number of misclassified points is minimized, is considerably more difficult and in fact is NP-complete, as shown in Proposition 2 of Section 2 below. This problem was considered in [8], where a parametric minimization approach was proposed and implemented in [1]. Although the parametric procedure is effective, it is costly computationally, which is to be expected since the underlying problem is NP-complete. In the present approach we shall propose a fast alternative hybrid criterion that is quite effective in approximately minimizing the number of misclassified points as determined by tenfold cross-validation [15]. The basic idea is to minimize the number of misclassified points by translating the separating plane, and then rotating the plane in order to minimize a weighted average sum of the distances of misclassified points to a separating plane. This hybrid separability criterion leads to an effective finite algorithm for solving the separation problem.

We outline the contents of the paper now. In Section 2 we define the misclassification minimization problem (7), and establish the NP-completeness of the equivalent problem (8) in Proposition 2. We then define our Hybrid Misclassification Minimization (HMM) Problem 3 and establish the existence of a global solution to it in Theorem 4, and prescribe a finite hybrid algorithm, HMM Algorithm 5, for its approximate solution. Section 3 contains numerical results that indicate that the proposed hybrid algorithm is fast and appears to generalize better than the parametric algorithm misclassification minimization [1].

A word about our notation now. For a vector x in the n-dimensional real space R^n, x_+ will denote the vector in R^n with components $(x_+)_i := \max\{x_i, 0\}$, $i = 1, \ldots, n$. Similarly x_* will denote the vector in R^n with components $(x_*)_i := (x_i)_*$, $i = 1, \ldots, n$, where $(\cdot)_*$ is the step function defined as one for positive x_i and zero otherwise. The norm $\|\cdot\|$ will denote the l_2 norm, while $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix, A^T will denote the transpose, and A_i will denote row i. For two vectors x and y in R^n, $x \perp y$ will denote $x^T y = 0$. A vector of ones in a real space of arbitrary dimension will be denoted by e. The notation $\arg\min\limits_{x \in S} f(x)$ will denote the set of minimizers of $f(x)$ on the set S. Similarly $\mathrm{argKKT}\min\limits_{x \in S} f(x)$ will denote the set of minimizers and corresponding Lagrange multipliers of the Karush-Kuhn-Tucker conditions for $\min\limits_{x \in S} f(x)$. By a separating plane, with respect to two given point sets $\mathcal{A}$ and $\mathcal{B}$ in R^n, we shall mean a plane that attempts to separate R^n into two half spaces such that each open halfspace contains points mostly of $\mathcal{A}$ or $\mathcal{B}$. The cardinality of a set $\mathcal{A}$ will be denoted by $\mathrm{card}(\mathcal{A})$. The symbol " := " defines a quantity appearing on its left by a quantity appearing on its right.

2 The Hybrid Misclassification Minimization Problem

We begin by defining the "pure" misclassification minimization problem as in [8] with the help of the step function $(\cdot)_*$. For the two finite point sets $\mathcal{A}$ and $\mathcal{B}$ in R^n, represented respectively by $A \in R^{m\times n}$ and $B \in R^{k\times n}$, we need to find a plane $x^Tw = \gamma$ such that as many as possible of the following inequalities are satisfied:

$$Aw > e\gamma,\ Bw < e\gamma. \tag{4}$$

Upon normalization, this is equivalent to satisfying as many as possible of the following inequalities

$$Aw \geq e\gamma + e,\ Bw \leq e\gamma - e. \tag{5}$$

Thus, we wish to minimize the number of misclassified points by the plane $x^Tw = \gamma$. This problem can be stated as the following misclassification minimization problem

$$\min_{w,\gamma} e^T(-Aw + e\gamma + e)_* + e^T(Bw - e\gamma + e)_*. \tag{6}$$

In [8, 1] this problem was reformulated as a linear program with equilibrium constraints (LPEC) [6], that is a linear program with a single complementarity constraint. An implicitly exact penalty method as well as a parametric method were proposed for solving the LPEC in [8] and successfully implemented in [1]. Although effective, the parametric approach is costly, because for each value of the parameter, a nonconvex bilinear program need to be solved. We propose here an alternative hybrid approach that is considerably faster and which appears to generalize better than the parametric approach.

The basic idea of the hybrid approach is to use two criteria for obtaining $(w, \gamma) \in R^{n+1}$ that characterizes the separating plane $x^Tw = \gamma$. More specifically, for a fixed γ we solve the linear program (3) to determine w. Then for this w we solve a one-dimensional minimization problem (6) in γ to minimize the number of misclassified points. The process is repeated until no improvement in the number of misclassified points is possible. We term such a point as a stationary point. The idea of using different criteria to determine different parts of the solution (w, γ) is similar to that of finding equilibrium points [13] and solving multicriteria optimization problems [14].

Before defining precisely our problem, we slightly modify the misclassification minimization problem (6) as follows:

$$\min_{w,\gamma} e^T(e - (Aw - e\gamma)_*) + e^T(e - (-Bw + e\gamma)_*) \tag{7}$$

We note that while (6) counts the number of violated normalized inequalities (5), the minimization problem (7) counts the number of violated un-normalized inequalities (4). In fact (7) is equivalent to maximizing the number of satisfied inequalities in (4), that is

$$\max_{w,\gamma} e^T(Aw - e\gamma)_* + e^T(-Bw + e\gamma)_* \tag{8}$$

We note further, as in the case of robust linear programming separation [2] achieved by the linear program (3), when the null solution $(w, \gamma) = (0, \pm 1) \in R^{n+1}$ gives a maximum value of $\max\{m, k\}$ for (8), then it is never unique in w. This is because any plane $w^Tx = \gamma$, with $w \neq 0$ will also achieve the same maximum by placing the appropriate set $\mathcal{A}$ or $\mathcal{B}$ in one of the open halfspaces it generates. This is a useful property of (8), otherwise the null solution $w = 0$ would pose a computational difficulty similar to that addressed in [2].

If we now assume that A and B have integer entries, then problem (8) belongs to the following class of problems which, we will show, is NP-complete.

1. Maximum Inequality Satisfiability (MIS). Let the matrix $H \in R^{p\times q}$ have integer entries. Find the maximum number of satisfiable inequalities

$$Hx > 0, \tag{9}$$

where x is a vector of q rational numbers.

Note that in (9), H plays the role of the matrix $\begin{bmatrix} A & -e \\ -B & e \end{bmatrix}$ of (4). We now show that this problem is NP-complete.

2. Proposition *The MIS Problem 1 is NP-complete.*

Proof The NP-complete Open Hemisphere (OH) Problem [4, page 246, problem MP6] is the problem of determining whether, for a positive integer $r \le p$, r of the inequalities $Hx > 0$ can be satisfied by a rational vector x.

We first show that MIS is in NP by reducing it to at most two instances of OH which is in NP. If we are given an integer $r \le p$, then we can decide whether it is a solution of MIS as follows. The integer r is a solution of MIS if and only if r is a solution of OH and $r+1$ is not a solution of OH when $r+1 \le p$. Since OH is in NP and checking whether r solves MIS can be performed by solving at most two instances of OH, it follows that MIS is in NP.

Now we show that MIS is NP-hard by reducing OH to an instance of MIS. Given a positive integer r, we solve MIS and obtain $\bar{r}$ for its maximum. The integer r solves OH if and only if $r \le \bar{r}$. $\square$

We note that Heath's NP-completeness result [5, Appendix C] is for a differently stated problem than ours. In particular, Heath minimizes

$$\min\{\text{card}\{i \mid A_i w > \gamma_i\}, \text{card}\{i \mid B_i w > \gamma_i\}\} + \min\{\text{card}\{i \mid A_i w < \gamma_i\}, \text{card}\{i \mid B_i w < \gamma_i\}\}$$

We believe that our measure of misclassification minimization as given in the (MIS) Problem 1 is simpler and more direct than Heath's.

In order to avoid the solution of the NP-complete problem (8), we replace it by the following problem, that is more tractable computationally: Translate the plane $x^T w = \gamma$ by varying γ so that it that maximizes the number of correctly classified points, where the plane orientation w has been determined by a rotation that minimizes a weighted average sum of distances of misclassified points to the plane. This results in the following hybrid misclassification minimization problem.

3. Hybrid Misclassification Minimization (HMM) Problem. Find $(\bar{w}, \bar{\gamma}) \in R^{n+1}$ that determine the plane $x^T \bar{w} = \bar{\gamma}$, such that $\bar{w} = \bar{w}(\bar{\gamma})$ and

$$\bar{\gamma} \in \arg\max_{\gamma} f(\gamma) := \arg\max_{\gamma} e^T(Aw(\gamma) - e\gamma)_* + e^T(-Bw(\gamma) + e\gamma)_* \tag{10a}$$

$$\text{such that} \;\; w(\gamma) \in \arg\min_{w} \frac{e^T}{m}(-Aw + e\gamma + e)_+ + \frac{e^T}{k}(Bw - e\gamma + e)_+ \tag{10b}$$

We note that for each $\gamma \in R$, the subproblem (10b) is equivalent to the linear program (3) with fixed γ. Because this linear program is feasible and its objective is bounded below by zero, it always has a solution. Hence the objective function $f(\gamma)$ of the HMM Problem (10a) is well defined. We show now that $f(\gamma)$ attains a maximum for some $\bar{\gamma} \in R$.

4. Theorem Existence of Solution to the HMM Problem 3 *For any $A \in R^{m\times n}, B \in R^{k\times n}$, the HMM Problem 3 has a solution $(\bar{w}, \bar{\gamma}) \in R^{n+1}$.*

Proof We observe first that $f(\gamma)$, as defined in (10a) is bounded above by $m+k$. Hence

$$\sup_{\gamma} f(\gamma) = \sigma < \infty.$$

Since $f(\gamma)$ takes on integer values only, it follows that σ is an integer. Hence, there exists a $\bar{\gamma}$ such that $f(\bar{\gamma}) > \sigma - \frac{1}{2}$, and hence $f(\bar{\gamma}) = \sigma$ and consequently

$$\bar{\gamma} \in \arg\max_{\gamma} f(\gamma).$$

$\square$

We state now our algorithm for solving the HMM Problem 3.

5. The Hybrid Misclassification Minimization (HMM) Algorithm.
Initialization

$$(w^0, \gamma^{-1}, y^0, z^0) \in \arg\min_{w,\gamma,y,z} \{\frac{e^T}{m}y + \frac{e^T}{k}z \mid Aw + y \geq e\gamma + e, Bw - z \leq e\gamma - e, y, z \geq 0\} \tag{11}$$

Iteration

$$\gamma^i \in \arg\max_{\gamma} g(w^i, \gamma) := \arg\max_{\gamma} e^T(Aw^i - e\gamma)_* + e^T(-Bw^i + e\gamma)_* \tag{12}$$

$$\text{Stop if } g(w^i, \gamma^i) \leq g(w^{i-1}, \gamma^{i-1}) \tag{13}$$

$$w^{i+1} \in \arg\min_{w} h(w, \gamma^i) := \arg\min_{w} \frac{e^T}{m}(-Aw + e\gamma^i + e)_+ + \frac{e^T}{k}(Bw - e\gamma^i + e)_+ \tag{14}$$

Note that the first subproblem (12) of the HMM algorithm is a one dimensional problem with a finite number of objective function values that lie in the set $\{0, 1, \cdots, m+k\}$ and is easily solved by a line search procedure. The second subproblem (14) is equivalent to the initialization linear program (11) with γ fixed at $\gamma = \gamma^i$.

Although the HMM Algorithm does not necessarily solve the HMM Problem 3, it does terminate very quickly after two to five iterations at a solution that is about as good as that obtained by the more complex parametric misclassification algorithm [8, 1]. Furthermore the HMM Algorithm appears to generalize better than the parametric algorithm, as indicated by the numerical computations given in the next section.

We state now a finite termination result for the HMM Algorithm 5.

6. Finite Termination of the HMM Algorithm The HMM Algorithm 5 terminates in a finite number of steps at a stationary point $(w^{\bar{i}-1}, \gamma^{\bar{i}-1})$ satisfying the stopping criterion (13).

Proof Since the sequence $\{g(w^i, \gamma^i)\}$ is in the finite set $\{0, 1, \cdots, m+k\}$, it cannot increase indefinitely. Hence at some iteration $\bar{i}$, it must satisfy the stopping criterion (13) and the HMM Algorithm 5 terminates. $\square$

We note that the stopping criterion (13) leads to a stationary point $(w^{\bar{i}-1}, \gamma^{\bar{i}-1})$ in the sense that $g(w^{\bar{i}-1}, \gamma^{\bar{i}-1}) = \max_{i\leq\bar{i}} \max_{\gamma} g(w^i, \gamma)$. In the real world problems solved in the next section, such a point seems to be as good as that obtained by a more complex and costly algorithm, and generalizes better.

3 Numerical Computation and Comparisons

We report now on numerical results on the Wisconsin Breast Cancer Database (WBCD) and other data sets from the Irvine Machine Learning Database Repository [10] as well as the Star/Galaxy database collected by Odewahn [12] and the Wisconsin Breast Cancer Prognosis Database [9, 16].

For each data set, a separating plane was obtained by three methods: the parametric misclassification minimization (PMM) procedure of [8, 1], the HMM Algorithm 5 of Section 2, and the robust linear program (RLP) algorithm [2], that is the linear program (3). In order to measure how well each separating plane generalizes to unseen data, we performed tenfold cross-validation on each data set [15]. Specifically, we divided each data set into ten equal parts, obtained a separating plane for the combined nine parts (training) and tested the correctness of the plane (generalization) on the tenth set. The percent generalization correctness for each data set was an average of the correctness over the ten different subsets used for training and testing. The time reported was the average time for the ten different subsets used for training.

The parametric misclassification minimization procedure was coded in the modeling language AMPL [3] in [1] utilizing the MINOS [11] linear programming solver. The HMM Algorithm and the robust linear program algorithm were implemented using C and called MINOS as a subroutine to solve the linear programs.

Table 1 gives a summary of the numerical results. To address the possibility that the reported CPU times might be biased against AMPL, because of the overhead involved when AMPL calls the MINOS solver, we have also included another comparative criterion: the average number of LPs solved by each method. We make the following additional observations.

(i) Testing set correctness:

HMM highest in 5 out of 10 cases
RLP highest in 3 out of 10 cases
PMM highest in 2 out of 10 cases

(ii) Training set correctness:

PMM highest in all 10 cases(as expected, since PMM maximizes this quantity)
HMM second highest in all 10 cases
RLP lowest in all 10 cases

(iii) Computing time:

RLP fastest all 10 cases. Total time 12.08 seconds (as expected, since it solves a single LP)
HMM same order of time as RLP. Total time 32.54 seconds
PMM slowest in all 10 cases. Total time 1600.18 seconds

(iv) Average of average number of LPs solved:

RLP constant of 1
HMM average of 2.32
PMM average of 22.3

Table 1. Comparison of Hybrid Misclassification Minimization (HMM) with Parametric Misclassification Minimization (PMM) [8, 1] & Robust Linear Programming (RLP) [2]

Date Set	m k n	Training Set Correctness Testing Set Correctness Time Seconds SPARCstation 20 Average LPs Solved		
		HMM	PMM	RLP
WBC Prognosis	28 119 32	89.12 72.24 0.71 2.1	95.92 71.33 10.65 13.8	84.343 66.048 0.501 1
WBCD	239 443 9	97.87 97.36 0.64 2	98.57 96.47 24.65 15.6	97.73 97.21 0.21 1
Cleveland Heart	216 81 14	87.50 82.84 0.41 2	91.43 82.16 17.67 26	84.47 83.51 0.22 1
Ionosphere	225 126 34	96.56 88.36 1.46 2.1	98.42 87.52 27.26 14	94.90 86.09 0.98 1
Liver Disorders	145 200 6	72.21 66.64 0.43 2.1	74.85 68.37 18.51 28.8	68.99 66.93 0.28 1
Pima Diabetes	268 500 8	78.42 75.89 1.51 4.4	80.55 76.67 51.40 40.4	76.77 76.00 0.75 1
Star/Galaxy(Dim)	2082 2110 14	95.98 95.63 19.73 2.1	96.52 95.42 1122.70 37.4	95.64 95.51 6.89 1
Star/Galaxy(Bright)	1505 957 14	99.68 99.23 4.95 2	99.89 99.19 266.13 8.1	99.62 99.39 0.87 1
Tic Tac Toe	626 332 9	68.93 66.16 2.11 2.2	69.12 64.50 46.45 26.9	62.75 60.23 1.16 1
Votes	168 267 16	98.03 95.62 0.59 2.2	98.82 94.01 14.76 12	97.45 95.63 0.22 1

4 Conclusion

We have introduced a fast hybrid misclassification minimization algorithm for minimizing the number of misclassified points by a plane attempting to separate two given sets in R^n. The algorithm essentially solves two to five linear programs to determine the orientation of the separating plane and translates the plane to minimize the number of misclassified points. The algorithm is simple and robust and appears to be a very promising tool for machine learning.

Acknowledgement

We are grateful to Kristin P. Bennett and Erin J. Bredensteiner for making available to us their AMPL program for the PMM algorithm for comparison purposes.

References

[1] K. P. Bennett and E. J. Bredensteiner. A parametric optimization method for machine learning. Department of Mathematical Sciences Report No. 217, Rensselaer Polytechnic Institute, Troy, NY 12180, 1994.

[2] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[3] R. Fourer, D. Gay, and B. Kernighan. *AMPL.* The Scientific Press, South San Francisco, California, 1993.

[4] M. R. Garey and D. S. Johnson. *Computers and Intractability, A Guide to the Theory of NP–Completeness.* W. H. Freeman and Company, San Francisco, 1979.

[5] David Heath. *A geometric Framework for Machine Learning.* PhD thesis, Department of Computer Science, Johns Hopkins University–Baltimore, Maryland, 1992.

[6] Z.-Q. Luo, J.-S. Pang, D. Ralph, and S.-Q. Wu. Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints. Technical Report 275, Communications Research Laboratory, McMaster University, Hamilton, Ontario, Hamilton, Ontario L8S 4K1, Canada, 1993. Mathematical Programming, to appear.

[7] O. L. Mangasarian. Multi-surface method of pattern separation. *IEEE Transactions on Information Theory*, IT-14:801–807, 1968.

[8] O. L. Mangasarian. Misclassification minimization. *Journal of Global Optimization*, 5:309–323, 1994.

[9] O. L. Mangasarian, W. Nick Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, July-August 1995.

[10] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, California, http://www.ics.uci.edu/AI/ML/MLDBRepository.html., 1992.

[11] B. A. Murtagh and M. A. Saunders. MINOS 5.0 user's guide. Technical Report SOL 83.20, Stanford University, December 1983. MINOS 5.4 Release Notes, December 1992.

[12] S. Odewahn, E. Stockwell, R. Pennington, R. Hummphreys, and W. Zumach. Automated star/galaxy discrimination with neural networks. *Astronomical Journal*, 103(1):318–331, 1992.

[13] H. E. Scarf. *The Computation of Economic Equilibria.* Yale University Press, New Haven, Conneticut, 1973.

[14] R. E. Steuer. *Multiple Criteria Optimization: Theory, Computation, and Application.* John Wiley and Sons, 1986.

[15] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147, 1974.

[16] W. H. Wolberg, W. N. Street, D. N. Heisey, and O. L. Mangasarian. Computer-derived nuclear grade and breast cancer prognosis. *Analytical and Quantitative Cytology and Histology.* To appear.