Knowledge-Based Linear Programming *

O. L. Mangasarian†

Abstract

We introduce a class of linear programs with constraints in the form of implications. Such linear programs arise in support vector machine classification, where in addition to explicit datasets to be classified, prior knowledge such as expert's experience in the form of logical implications, are imposed on the classifier. The overall problem can be viewed either as a semi-infinite linear program or as a linear program with equilibrium constraints which, in either case, can be solved by an equivalent simple linear program under mild assumptions.

1 Introduction

Recently support vector machines (SVMs) [15, 3, 16, 14, 4, 13], which constitute the method of choice for classification and other approximation problems, were applied to problems where prior expert knowledge, such as doctors' experience, was incorporated into the classification problem in order to improve the correctness of the classifier generated [5, 6]. This resulted in a linear program where constraints corresponding to prior knowledge were made up of logical implications. In this work, we generalize this type of constraint and call the corresponding linear program *a knowledge-based linear program*. It is interesting to note that such knowledge-based logical implications are equivalent to an infinite number of constraints, thus resulting in a semi-infinite linear program [7], or alternatively they can be represented as constraints in a linear program with equilibrium constraints [8]. However, we will show that in Section 3 that such logical implications can be be replaced by a finite number of inequality constraints, thus leading to a simple linear programming formulation. In Section 2 of the paper we state our general knowledge-based linear program (KBLP) and motivate it by a knowledge-based linear support vector machine classifier. In Section 4 we establish equivalence of KBLP to a single linear program, give some duality results, and establish equivalence of KBLP to a linear program with equilibrium constraints. Section 5 concludes the paper with a summary and an outlook.

*Data Mining Institute Technical Report 03-04, July 2003. Research supported by NSF GrantCR-0138308 and by Microsoft.

† *olvi@cs.wisc.edu.* Computer Sciences Department, University of Wisconsin, Madison, WI 53706, and Department of Mathematics, University of California at San Diego, La Jolla, CA 92093.

A word about our notation. All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. The scalar product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$. For $x \in R^n$ and $p \in [1, \infty)$, the norm $\|x\|_p$ will denote the p-norm: $(\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ and $\|x\|_\infty$ will denote $\max_{1 \le i \le n} |x_i|$. For $x \in R^n$, $(x_+)_i = \max\{0, x_i\}$, $i = 1, \ldots, n$, while the sign function $sign(x)$ is defined as $sign(x)_i = 1$ if $x_i > 0$ else $sign(x)_i = -1$ if $x_i \le 0$, for $i = 1, \ldots, n$. For $x \in R^n$ and $y \in R^n$, $x \perp y$ denotes orthogonality, that is $x'y = 0$. For an $m \times n$ matrix A, A_i will denote the ith row of A, $A_{\cdot j}$ will denote the jth column of A and A_{ij} will denote the element in row i and column j. The identity matrix in a real space of arbitrary dimension will be denoted by I, while a column vector of ones of arbitrary dimension will be denoted by e.

2 The Knowledge-Based Linear Program (KBLP)

We consider the following linear program with a logical implication constraint:

$$\begin{array}{rrcl} \min\limits_{x} & c'x & & \\ s.t. & Hx & \ge & h, \\ & Ls \le d & \Longrightarrow & (Ex - a)'s \ge \mu. \end{array} \tag{1}$$

Here, $H \in R^{m \times n}$, $L \in R^{\ell \times k}$ and $E \in R^{k \times n}$, are given matrices, $c \in R^n$, $h \in R^m$, $d \in R^\ell$ and $a \in R^k$, are given vectors, and μ is a given real number. An alternate way to write this problem is the following semi-infinite program:

$$\begin{array}{rrcl} \min\limits_{x} & c'x & & \\ s.t. & Hx & \ge & h, \\ & (Ex - a)'s \ge \mu & & \forall : s \in \{s | Ls \le d\}. \end{array} \tag{2}$$

We describe now knowledge-based classification problems [5] that lead to the knowledge-based linear program (1)-(2). These problems, in addition to classifying conventional datasets consisting of a finite number of points in R^n, attempt to classify polyhedral sets in R^n which are known to belong to one of two classes. We thus consider the problem of classifying m points in the n-dimensional input space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the class $A+$ or $A-$ as specified by a given $m \times m$ diagonal matrix D with plus ones or minus ones along its diagonal. For this problem, the linear programming support vector machine [11, 1] with a linear kernel, which is a variant of the standard support vector machine [16, 2], is given by the following linear program with parameter $\nu > 0$:

$$\begin{array}{rrcl} \min\limits_{(w,\gamma,y,t) \in R^{n+1+m}} & \nu e'y + e't & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \ge & e \\ & t \ \ge w & \ge & -t \\ & y & \ge & 0. \end{array} \tag{3}$$

Here, w is the normal to the bounding planes:

$$\begin{array}{rcl} x'w & = & \gamma + 1 \\ x'w & = & \gamma - 1, \end{array} \tag{4}$$

that bound most of the points belonging to the sets $A+$ and $A-$ respectively. The constant γ determines their location relative to the origin. When the two classes are strictly linearly separable, that is when the error variable $y = 0$ in (3), the plane $x'w = \gamma + 1$ bounds all of the class $A+$ points, while the plane $x'w = \gamma - 1$ bounds all of the class $A-$ points as follows:

$$\begin{array}{llll} A_i w & \geq \gamma + 1, & \text{for} & D_{ii} = 1 \\ A_i w & \leq \gamma - 1, & \text{for} & D_{ii} = -1. \end{array} \tag{5}$$

Consequently, the plane:

$$x'w = \gamma, \tag{6}$$

midway between the bounding planes (4), is a separating plane that separates points belonging to $A+$ from those belonging to $A-$ completely if $y = 0$, else only approximately. The 1-norm term $\|w\|_1$ in (3), that is the term $e't$ at a solution point, which is half the reciprocal of the distance $\frac{2}{\|w\|_1}$ measured using the ∞-norm distance [10] between the two bounding planes of (4), maximizes this distance, often called the "margin". Maximizing the margin enhances the generalization capability of a support vector machine [16, 2]. If the two classes are linearly inseparable, then the two planes bound the two classes with a "soft margin" (i.e. bound approximately with some error) determined by the nonnegative error variable y, that is:

$$\begin{array}{lllll} A_i w & + & y_i & \geq & \gamma + 1, \ \text{for } D_{ii} = 1 \\ A_i w & - & y_i & \leq & \gamma - 1, \ \text{for } D_{ii} = -1. \end{array} \tag{7}$$

The 1-norm of the error variable y is minimized parametrically with weight ν in (3), resulting in an approximate separating plane. This plane acts as a linear classifier as follows:

$$sign(x'w - \gamma\ ) \begin{cases} = 1, & \text{then } x \in A+ \\ = -1, & \text{then } x \in A-, \end{cases} \tag{8}$$

where $sign(\cdot)$ is the sign function defined in the Introduction. Suppose now that we have prior information of the following type. All points $x \in R^n$ lying in the polyhedral set determined by the linear inequalities:

$$Bx \leq b, \tag{9}$$

where $B \in R^{k \times n}$ and $b \in R^k$, belong to class $A+$. Such inequalities generalize simple box constraints such as $a \leq x \leq d$. Looking at the inequalities (9) we conclude that the following implication must hold:

$$Bx \leq b \implies x'w \geq \gamma + 1. \tag{10}$$

That is, the *knowledge set* $\{x \mid Bx \leq b\}$ lies on the $A+$ side of the bounding plane $x'w = \gamma + 1$. This leads to the following knowledge-based support vector machine classification problem:

$$\begin{array}{rrcl}
\displaystyle\min_{(w,\gamma,y,t)\in R^{n+1+m}} & \nu e'y + e't & & \\
\text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\
& t \ \geq w & \geq & -t \\
& y & \geq & 0 \\
& Bs \leq b & \Longrightarrow & s'w \geq \gamma + 1,
\end{array} \tag{11}$$

which is a special case of the knowledge-based linear program (1) upon making the appropriate identifications for $x, c, H, h, L, d, E, a, \mu$ in (1) in terms of $w, \gamma, y, t, \nu, D, A, B, b$ of (11).

We show now that the implication constraint (10) is equivalent to a finite number of inequality constraints and hence KBLP (1) is equivalent to a simple linear program.

3 Equivalence of KBLP to a Linear Program with a Finite number of Linear Inequalities

The equivalence result can be obtained as a direct consequence of the set containment characterizations on [12, Proposition 2.1] or more directly by Motzkin's Theorem of the Alternative [9, Theorem 2.4.2] which we utilize here.

Proposition 3.1 Implication Constraint as Linear Inequalities *Let the set $\{s \,|\, Ls \leq d\}$ be nonempty. Then, for for a fixed $x \in R^n$, the implication constraint:*

$$Ls \leq d \implies (Ex - a)'s \geq \mu, \tag{12}$$

is equivalent to the following system of linear inequalities having a solution $u \in R^\ell$:

$$\begin{array}{rlcl}
L'u+ & Ex & = & a \\
-d'u & & \geq & \mu \\
u & & \geq & 0.
\end{array} \tag{13}$$

Proof For a fixed $x \in R^n$, the implication (12) is equivalent to the following system having *no* solution $(s, \zeta) \in R^{k+1}$:

$$Ls - d\zeta \leq 0,\ (Ex - a)'s - \mu\zeta < 0,\ -\zeta < 0. \tag{14}$$

By Motzkin's Theorem [9, Theorem 2.4.2] this is equivalent to the following system *having* a solution $(u, \xi, \eta) \in R^{\ell+2}$:

$$L'u + (Ex - a)\xi = 0,\ -d'u - \mu\xi - \eta = 0,\ (u, \xi, \eta) \geq 0,\ (\xi, \eta) \neq 0. \tag{15}$$

If $\xi = 0$, then $\eta > 0$ and the system (15) having such a solution contradicts the nonemptiness assumption on $\{s \,|\, Ls \le d\}$, because for $s \in \{s \,|\, Ls \le d\}$ we obtain the contradiction:

$$0 = s'L'u \le u'd = -\eta < 0. \tag{16}$$

Hence the system (15) must have a solution with $\xi > 0$. Dividing throughout by ξ, and redefining $\frac{u}{\xi} \to u$, $\frac{\eta}{\xi} \to \eta$, the system (15) becomes:

$$L'u + (Ex - a) = 0,\ -d'u - \mu = \eta \ge 0,\ u \ge 0, \tag{17}$$

which is equivalent to (13).$\square$

With the above proposition the following equivalence becomes evident.

Corollary 3.2 Equivalent KBLP *Under the assumption that the set $\{s \,|\, Ls \le d\}$ is nonempty, the KBLP (1) is equivalent to the following linear program with a finite number of linear inequalities:*

$$\begin{array}{llcrcl} \min\limits_{x,u} & c'x & & & & \\ s.t. & Hx & & & \ge & h \\ & Ex+ & & L'u & = & a \\ & & & -d'u & \ge & \mu \\ & & & u & \ge & 0. \end{array} \tag{18}$$

We turn now to establishing existence of solution to KBLP (1) through the above equivalence and to some duality results.

4 Existence of Solution and Duality for KBLP

We will first establish that KBLP (1) has a solution provided that it is feasible and that the underlying linear program without the knowledge constraint is solvable.

Proposition 4.1 Existence of Solution to KBLP *Let KBLP (1) be feasible, that is:*

$$X = \{x \,|\, Hx \ge h,\ (Ex - a)'s \ge \mu,\ \forall s:\ Ls \le d\} \ne \emptyset, \tag{19}$$

and that the underlying linear program:

$$\begin{array}{llcl} \min\limits_{x} & c'x & & \\ s.t. & Hx & \ge & h, \end{array} \tag{20}$$

is solvable, then KBLP (1) is solvable.

Proof The dual of the underlying linear program (20) is:

$$\begin{array}{llcl} \max\limits_{t} & h't & & \\ s.t. & H't & = & c, \\ & t \ge & 0, & \end{array} \tag{21}$$

while the dual of the equivalent KBLP (18) is:

$$\begin{array}{rllllll} \max\limits_{v,r,\zeta} & h'v+ & a'r+ & \mu\zeta & & \\ s.t. & H'v+ & E'r & & = & c \\ & & Lr- & d\zeta & \leq & 0 \\ & (v, & & \zeta) & \geq & 0. \end{array} \tag{22}$$

The assumption that KBLP (1) is feasible, that is $X \neq \emptyset$, is equivalent to the fact the equivalent KBLP (18) is feasible. Since the underlying linear program (20) is solvable, its dual (21) is feasible. It follows that the dual (22) of the equivalent KBLP (18) is also feasible by taking $v = t$, $r = 0$, $\zeta = 0$. Hence the equivalent KBLP (18) is solvable because it is feasible and its dual (22) is feasible. Hence KBLP (1) is solvable. $\square$

Having established existence of solution to KBLP (1), we pose the question: When is the dual (22), to the equivalent KBLP (18), also a KBLP? That is, can the constraints of (22) be reduced to logical implication constraints and linear inequalities? In its present general form, the dual problem does not seem to be amenable to such a reduction. However if we take a special case of $h = 0$, $d = 0$, $\mu = 0$, then indeed both the primal KBLP (1) and its dual (22) are KBLPs as follows.

Proposition 4.2 Dual KBLPs *Let*

$$Y = \{x \,|\, Hx \geq 0,\, (Ex - a)'s \geq 0,\, \forall s: \, Ls \leq 0\} \neq \emptyset. \tag{23}$$

Then the following problems are dual knowledge-based linear programs for which standard linear programming duality results hold:

$$\begin{array}{rrcl} \min\limits_{x} & c'x & & \\ s.t. & Hx & \geq & 0 \\ & Ls \leq 0 & \Longrightarrow & (Ex - a)'s \geq 0. \end{array} \tag{24}$$

$$\begin{array}{rrcl} \max\limits_{r} & a'r & & \\ s.t. & Lr & \leq & 0 \\ & Hy \leq 0 & \Longrightarrow & (-E'r + c)'y \leq 0. \end{array} \tag{25}$$

Proof Since (24) is a special case of KBLP (1), its dual is given as a special case of (22), the dual of KBLP (1): as:

$$\begin{array}{rlllll} \max\limits_{v,r,\zeta} & & a'r & & \\ s.t. & H'v+ & E'r & = & c \\ & & Lr & \leq & 0 \\ & v & & \geq & 0. \end{array} \tag{26}$$

For a fixed $r \in R^k$ it follows, by the Farkas Theorem [9, Theorm 2.4.6], that the first and last constraints of (26) are equivalent to the following system *not* having a solution $y \in R^n$:

$$Hy \leq 0,\, (c - E'r)'y > 0, \tag{27}$$

which in turn is equivalent to the implication:

$$Hy \leq 0, \implies, \ (-E'r + c)'y \leq 0. \tag{28}$$

Hence if we replace the first and last constraints of (26) by the equivalent implication of (28), we obtain the desired KBLP dual (25). $\Box$

One final interesting result is a connection to linear programs with equilibrium constraints (LPEQs) [8]. We will show that our KBLP (1) can be rewritten as an LPEQ that can be solved as a single linear program (18). To set the KBLP as an LPEQ note that KBLP (1) can be written in the equivalent form:

$$\min_x \{c'x \mid Hx \geq h, \ \min_s \{(Ex - a)'s \mid Ls \leq d\} \geq \mu\}. \tag{29}$$

The dual to the inner linear program in (29) is given by:

$$\max_u - d'u \text{ s.t. } - L'u = Ex - a, \ u \geq 0. \tag{30}$$

Hence the formulation (29) can be rewritten, using the KKT conditions for the inner linear program as the following LPEQ:

$$\min_{x,s,u} \{c'x \mid Hx \geq h, \ Ex + L'u = a, \ d'u + \mu \leq 0, 0 \leq u \perp (-Ls + d) \geq 0\}. \tag{31}$$

However, as we have established in Corollary 3.2, that the orthogonality condition $u \perp (-Ls + d) \geq 0$ is not needed and can be dropped. So the LPEQ (31) can be solved as the single linear program (18). A possible intuitive justification for this is that primal and dual feasibility for the inner linear program of (29) is sufficient, without the orthogonality condition $u \perp (-Lu + d) \geq 0$, to induce the condition $(Ex - a)'s \geq -d'u \geq \mu$ for any $s \in R^k$ such that $Ls \leq d$, which is the implication constraint of KBLP (1).

5 Conclusion

We have established the equivalence of linear programs with logical implication constraints to conventional linear programs with a finite number of inequalities. We refer to such linear programs as knowledge-based linear programs (KBLPs) because they arise naturally in classification problems where these implication constraints are generated by prior knowledge. Since KBLPs can also be written as semi-infinite linear programs or as linear programs with equilibrium constraints, it follows from our equivalence result that such semi-infinite linear programs and linear programs with equilibrium constraints can be solved as single linear programs with a finite number of constraints. KBLPs can be generalized in a number of ways including the handling implication constraints involving nonlinear terms in the primal variable x. One such generalization using nonlinear kernels has been proposed in [6]. Another interesting application is to approximation problems where prior knowledge can be used to obtain better approximations. This avenue of research is currently being investigated. It would be interesting to find other real world applications that can be modeled as knowledge-based linear or convex programs.

References

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[2] V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods*. John Wiley & Sons, New York, 1998.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, MA, 2000.

[4] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 171–203, Cambridge, MA, 2000. MIT Press.

[5] G. Fung, O. L. Mangasarian, and J. Shavlik. Knowledge-based support vector machine classifiers. Technical Report 01-09, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, November 2001. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-09.ps, NIPS 2002 Proceedings, to appear.

[6] G. Fung, O. L. Mangasarian, and J. Shavlik. Knowledge-based nonlinear kernel classifiers. Technical Report 03-02, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, March 2003. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/02-03.ps. Conference on Learning Theory (COLT 03) and Workshop on Kernel Machines, Washington D.C., August 24-27, 2003, to appear.

[7] M. A. Goberna and M. A. López. *Linear Semi-Infinite Optimization*. John Wiley, New York, 1998.

[8] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, England, 1996.

[9] O. L. Mangasarian. *Nonlinear Programming*. SIAM, Philadelphia, PA, 1994.

[10] O. L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

[11] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[12] O. L. Mangasarian. Set containment characterization. Technical Report 01-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, November 2001. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-10.ps. *Journal of Global Optimization* 24, 2002, 473-480.

[13] B. Schölkopf, C. Burges, and A. Smola (editors). *Advances in Kernel Methods: Support Vector Machines.* MIT Press, Cambridge, MA, 1999.

[14] A. Smola, P. L. Bartlett, B. Schölkopf, and J. Schuurmann (editors). *Advances in Large Margin Classifiers.* MIT Press, Cambridge, MA, 2000.

[15] A. Smola and B. Schölkopf. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[16] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, second edition, 2000.