Centre de Recherches Mathématiques
CRM Proceedings and Lecture Notes

Nonlinear Knowledge in Kernel Machines

Olvi L. Mangasarian and Edward W. Wild

Abstract. We give a unified presentation of recent work in applying prior knowledge to nonlinear kernel approximation [MW05] and nonlinear kernel classification [MW06]. In both approaches, prior knowledge over general nonlinear sets is incorporated into nonlinear kernel approximation or classification problems as linear constraints in a linear program. The key tool in this incorporation is a theorem of the alternative for convex functions that converts nonlinear prior knowledge implications into linear inequalities without the need to kernelize these implications. Effectiveness of the proposed approximation formulation is demonstrated on two synthetic examples as well as an important lymph node metastasis prediction problem arising in breast cancer prognosis. Effectiveness of the proposed classification formulation is demonstrated on three publicly available datasets, including a breast cancer prognosis dataset. All these problems exhibit marked improvements upon the introduction of prior knowledge of nonlinear kernel approximation and classification approaches that do not utilize such knowledge.

1. Introduction

Prior knowledge has been used effectively in improving classification both for linear [FMS03b] and nonlinear [FMS03a] kernel classifiers as well as for nonlinear kernel approximation [MSW04, MST+05]. In all these applications prior knowledge was converted to linear inequalities that were imposed on a linear program. The linear program generated a linear or nonlinear classifier, or a linear or nonlinear function approximation, all of which were more accurate than the corresponding results that did not utilize prior knowledge. However, whenever a nonlinear kernel was utilized in these applications, kernelization of the prior knowledge was not a transparent procedure that could be easily related to the original sets over which prior knowledge was given. In contrast, prior knowledge over arbitrary general sets has been recently incorporated *without* kernelization of the prior knowledge sets into nonlinear kernel approximation [MW05] and nonlinear kernel classification [MW06]. Here, we present a unified formulation of both approaches, which is made possible through the use of a fundamental theorem of the alternative for convex functions that we describe in Section 2 of the paper. An interesting, novel approach to knowledge-based support vector machines that modifies the hypothesis space rather

Key words and phrases. prior knowledge, kernel classification, kernel approximation, linear programming, theorem of the alternative.

than the optimization problem is given in [LSG06]. In another recent approach, prior knowledge is incorporated by adding additional points labeled based on the prior knowledge to the dataset [MSWT06].

In Section 2 we describe a general formulation for incorporating prior knowledge into kernel machines. Section 3 shows how this formulation can be specialized to incorporate prior knowledge into nonlinear kernel approximation, while Section 4 shows how to specialize the formulation to incorporate prior knowledge into nonlinear kernel classification. Numerical examples from show prior knowledge can improve both approximation and classification are presented in Section 5. Section 6 concludes the paper.

We describe our notation now. All vectors will be column vectors unless transposed to a row vector by a prime $'$. The scalar (inner) product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$. For $x \in R^n$,$\|x\|_1$ denotes the 1-norm: $\sum_{i=1}^{n} |x_i|)$ while $\|x\|$ denotes the 2-norm: $\sum_{i=1}^{n}(x_i)^2)^{\frac{1}{2}}$ and x_+ denotes the vector $\max\{x, 0\}$. The notation $A \in R^{m\times n}$ will signify a real $m \times n$ matrix. For such a matrix, A' will denote the transpose of A, A_i will denote the i-th row of A and $A_{\cdot j}$ the j-th column of A. A vector of ones in a real space of arbitrary dimension will be denoted by e. Thus for $e \in R^m$ and $y \in R^m$ the notation $e'y$ will denote the sum of the components of y. A vector of zeros in a real space of arbitrary dimension will be denoted by 0. For $A \in R^{m\times n}$ and $B \in R^{n\times k}$, a *kernel* $K(A, B)$ maps $R^{m\times n} \times R^{n\times k}$ into $R^{m\times k}$. In particular, if x and y are column vectors in R^n then, $K(x', y)$ is a real number, $K(x', B')$ is a row vector in R^m and $K(A, B')$ is an $m \times m$ matrix. We shall make no assumptions whatsoever on our kernels other than symmetry, that is $K(x', y)' = K(y', x)$, and in particular we shall not assume or make use of Mercer's positive definiteness condition [Vap00, SS02, CST00]. The base of the natural logarithm will be denoted by ε. A frequently used kernel in nonlinear classification is the Gaussian kernel [Vap00, CM98, Man00] whose ij-th element, $i = 1, \ldots, m,\ j = 1, \ldots, k$, is given by: $(K(A, B))_{ij} = \varepsilon^{-\mu\|A_i{}' - B_{\cdot j}\|^2}$, where $A \in R^{m\times n}$, $B \in R^{n\times k}$ and μ is a positive constant. The abbreviation "s.t." stands for "subject to".

2. General Formulation

We wish to impart knowledge to a function learned on a dataset in R^n represented by the m rows of the matrix $A \in R^{m\times n}$ with associated labels represented by the m-dimensional vector d. For approximation problems, d will be real valued, that is $d \in R^m$. For classification problems, d will have values $+1$ or -1 according to the class of each example, that is $d \in \{-1, +1\}^m$. The learned function f from R^n to R is defined as follows:

$$f(x) = K(x', B')u - \gamma \tag{1}$$

where $B \in R^{k\times n}$ is an arbitrary basis matrix and $K(x', B') : R^{1\times n} \times R^{n\times k} \longrightarrow R^{1\times k}$ is an arbitrary kernel function. In an approximation setting, we will use $f(x)$ directly to obtain a predicted value at $x \in R^n$, while in a classification setting we will use the value of the sign of $f(x)$ to obtain the predicted class of x. In general, the matrix B is set equal to A [Man00]. However, in reduced support vector machines [LM01, HL04] $B = \bar{A}$, where $\bar{A}$ is a submatrix of A whose rows are a small subset of the rows of A. In fact, B can be an arbitrary matrix in $R^{k\times n}$. The variables $u \in R^k$

and $\gamma \in R$ are parameters determined by the following optimization problem:

$$\min_{u,\gamma} \ \nu L(u, \gamma, A, d) + \|u\|_1, \tag{2}$$

where L is an arbitrary loss function, and ν is a positive parameter that controls the weight in the optimization problem of data fitting versus complexity reduction. In approximation problems, the loss function L will measure the difference between the actual values of the given data d and the predicted values given by f. For classification problems, the loss function L will be related to misclassification rate. The 1-norm of u is chosen in order to promote solution sparsity [BM98, ZRHT04]. In Section 3 we will demonstrate that the use of the 1-norm of the error between the predicted and actual values for L in an approximation setting will result in a linear program, while Section 4 shows that use of the hinge loss for L results in a linear program for classification problems.

We now impose prior knowledge on the construction of our learned function $f(x) = K(x', B')u - \gamma$ through the following implication:

$$g(x) \leq 0 \Longrightarrow K(x', B')u - \gamma \geq \phi(x), \ \forall x \in \Gamma. \tag{3}$$

Here, $g(x) : \Gamma \subset R^n \longrightarrow R^p$ is a p-dimensional function defined on a subset Γ of R^n that determines the region in the input space where prior knowledge requires that $K(x', B')u - \gamma$ be larger than some known function $\phi(x) : \Gamma \subset R^n \longrightarrow R$. In previous work [MSW04, FMS02, FMS03a], prior knowledge implications such as (3) could not be handled directly as we shall do here by using Theorem 2.1 below. Instead, the inequality $g(x) \leq 0$ was kernelized. This led to an inequality not easily related to the original constraint $g(x) \leq 0$. In addition, in [MSW04, FMS02, FMS03a] could only handle linear $g(x)$ and $\phi(x)$. The implication (3) can be written in the following equivalent logical form:

$$\begin{array}{c} g(x) \leq 0, \ K(x', B')u - \gamma - \phi(x) < 0, \\ \text{has no solution } x \in \Gamma. \end{array} \tag{4}$$

It is precisely implication (3) that we shall convert to a system which is linear in the parameters of f, (u, γ), by means of the following theorem of the alternative for convex functions. The alternatives here are that either the negation of (4) holds, or (5) below holds, but *never both*.

Theorem 2.1. Prior Knowledge as System of Linear Inequalities *For a fixed $u \in R^k, \gamma \in R$, the following are equivalent:*

(i) *The implication (3) or equivalently (4) holds.*

(ii) *There exists $v \in R^p$, $v \geq 0$ such that:*

$$K(x', B')u - \gamma - \phi(x) + v'g(x) \geq 0, \ \forall x \in \Gamma, \tag{5}$$

where it is assumed for the implication (i)$\Longrightarrow$(ii) **only***, that $g(x)$ and $K(x', B')$ are convex on Γ, $\phi(x)$ is concave on Γ, Γ is a convex subset of R^n, $u \geq 0$ and that $g(x) < 0$ for some $x \in \Gamma$.*

Proof (i)$\Longrightarrow$(ii): This follows from [Man69, Corollary 4.2.2], the fact that the functions $g(x)$ and $K(x', B')u - \gamma - \phi(x)$ of (4) are convex on Γ and that $g(x) < 0$ for some $x \in \Gamma$.

(i)$\Longleftarrow$(ii): If (i) did not hold then there there exists an $x \in \Gamma$ such that $g(x) \leq 0, \ K(x', B')u - \gamma - \phi(x) < 0$, which would result in the contradiction:

$$0 > K(x', B')u - \gamma - \phi(x) + v'g(x) \geq 0. \square \tag{6}$$

We note immediately that in the proposed application of converting prior knowledge to linear inequalities in the parameters (u, γ) all we need is the implication (i)$\Longleftarrow$(ii), which **requires no assumptions whatsoever** on the functions $g(x)$, $K(x', B')$, $\phi(x)$ or on the parameter u. We further note that the implication (3) can represent fairly complex knowledge such as $K(x', B')u - \gamma$ being equal to any desired function whenever $g(x) \leq 0$.

We note that Theorem 2.1 can also be invoked on the following prior knowledge implication, which is similar to (3):

$$h(x) \leq 0 \Longrightarrow K(x', B')u - \gamma \leq -\psi(x), \; \forall x \in \Lambda. \tag{7}$$

We now incorporate the prior knowledge contained in implications (3) and (7) into the optimization problem (2) as follows:

$$\begin{array}{rl} \min\limits_{u,\gamma,z_1,\ldots,z_\ell,q_1,\ldots,q_t} & \nu L(u, \gamma, A, d) + \|u\|_1 + \sigma(\sum\limits_{i=1}^{\ell} z_i + \sum\limits_{j=1}^{t} q_j) \\ \text{s.t.} & K({x^i}', B')u - \gamma - \phi(x^i) + v'g(x^i) + z_i \geq 0, \\ & z_i \geq 0, \; i = 1, \ldots, \ell, \\ & v \geq 0, \\ & -K({x^j}', B')u + \gamma - \psi(x^j) + r'h(x^j) + q_j \geq 0, \\ & q_j \geq 0, \; j = 1, \ldots, t, \\ & r \geq 0. \end{array} \tag{8}$$

We note that we have discretized the variable $x \in \Gamma$ and $x \in \Lambda$ in the constraints above to the finite meshes of points $\{x^1, x^2, \ldots, x^\ell\}$ and $\{x^1, x^2, \ldots, x^t\}$ in order to convert a semi-infinite program [GL98] with an infinite number of constraints into a finite mathematical program. We have also added nonnegative slack variables $z_1, z_2, \ldots, z_\ell$ and $q_1, q_2, \ldots, q_t$ to allow small deviations in the prior knowledge. The sum of these nonnegative slack variables for the prior knowledge inequalities is minimized with weight $\sigma > 0$ in the objective function in order to drive them to zero to the extent possible. Thus, the magnitude of the parameter σ enforces prior knowledge while the magnitude of ν enforces data fitting.

We turn now to specific formulations for knowledge-based kernel approximation and knowledge-based kernel classification.

3. Knowledge-Based Kernel Approximation

In the approximation setting, we wish to approximate a function, f, given exact or approximate function values of a dataset of points represented by the rows of the matrix $A \in R^{m \times n}$. Thus, for each point A_i we are given an exact or inexact value of f, denoted by a real number d_i, $i = 1, \ldots, m$. We therefore desire the parameters (u, γ) of (1) to be determined such that:

$$K(A, B')u - e\gamma - d \approx 0. \tag{9}$$

An appropriate choice of L to enforce the above condition is:

$$L(u, \gamma, A, d) = \|K(A, B')u - e\gamma - d\|_1. \tag{10}$$

This loss function is the sum of the absolute values of the differences between the predicted values $f(A_i)$ and the given values d_i, $i = 1, \ldots, m$. In order to incorporate this loss function into the optimization problem (8) we introduce a vector $s \in R^m$ defined by:

$$-s \leq K(A, B')u - e\gamma - d \leq s. \tag{11}$$

We can then incorporate the loss defined by (10) into (8) as follows:

$$
\begin{array}{lrl}
 & \displaystyle\min_{u,\gamma,z_1,\ldots,z_\ell,q_1,\ldots,q_t,s} & \nu\|s\|_1 + \|u\|_1 + \sigma(\sum_{i=1}^{\ell} z_i + \sum_{j=1}^{t} q_j) \\
 & \text{s.t.} & K(x^{i\prime}, B')u - \gamma - \phi(x^i) + v'g(x^i) + z_i \geq 0, \\
 & & z_i \geq 0,\ i = 1, \ldots, \ell, \\
(12) & & v \geq 0, \\
 & & -K(x^{j\prime}, B')u + \gamma - \psi(x^j) + r'h(x^j) + q_j \geq 0, \\
 & & q_j \geq 0,\ j = 1, \ldots, t, \\
 & & r \geq 0, \\
 & & -s \leq K(A, B')u - \gamma - d \leq s.
\end{array}
$$

Note that at the solution of (12), $\|s\|_1 = \|K(A, B')u - \gamma - d\|_1$. Note further that (12) is the same as [MW05, Equation 10], except that here we have explicitly included implication (7) in addition to (3).

4. Knowledge-Based Kernel Classification

The classification problem consists of classifying the points represented by the rows of the matrix $A \in R^{m\times n}$ into the two classes $+1$ and -1 according to their labels given as $d \in \{-1, +1\}^m$. Thus, for each point A_i we are given its label $d_i \in \{-1, +1\}$, $i = 1, \ldots, m$ and we seek to find a function f of the form in (1) that satisfies, to the extent possible, the separation condition:

$$
D(K(A, B')u - e\gamma) \geq 0, \tag{13}
$$

where $D \in R^{m\times m}$ is the diagonal matrix with diagonal d. Note that this condition is satisfied if and only if d_i and $f(A_i)$ both have the same sign. Support vector machines attempt to impose a stronger condition, $D(K(A, B')u - e\gamma) \geq e$, using the hinge loss:

$$
L(u, \gamma, A, d) = \|(e - D(K(A, B')u - e\gamma))_+\|_1. \tag{14}
$$

Note that the hinge loss involves the plus function, $(x)_+ = max\{x, 0\}$, introduced in Section 1. This loss function can be added to the optimization problem (8) by introducing a nonnegative slack variable $s \in R^m$ as follows:

$$
D(K(A, B')u - e\gamma) + s \geq e, s \geq 0. \tag{15}
$$

The loss (14) is incorporated into the optimization problem (8) as follows:

$$
\begin{array}{lrl}
 & \displaystyle\min_{u,\gamma,z_1,\ldots,z_\ell,q_1,\ldots,q_t,s} & \nu\|s\|_1 + \|u\|_1 + \sigma(\sum_{i=1}^{\ell} z_i + \sum_{j=1}^{t} q_j) \\
 & \text{s.t.} & K(x^{i\prime}, B')u - \gamma - \phi(x^i) + v'g(x^i) + z_i \geq 0, \\
 & & z_i \geq 0,\ i = 1, \ldots, \ell, \\
 & & v \geq 0, \\
(16) & & -K(x^{j\prime}, B')u + \gamma - \psi(x^j) + r'h(x^j) + q_j \geq 0, \\
 & & q_j \geq 0,\ j = 1, \ldots, t, \\
 & & r \geq 0, \\
 & & D(K(A, B')u - e\gamma) + s \geq 0, \\
 & & s \geq 0.
\end{array}
$$

We note that at the solution, $\|s\|_1 = \|(e-D(K(A, B')u-e\gamma))_+\|_1$. Note further that (16) is the same as the optimization problem in [MW06] with $\phi(x) = \psi(x) = \alpha,\ \forall x$.

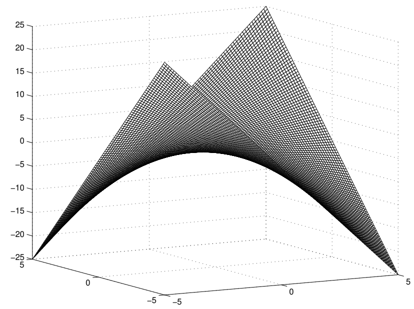

FIGURE 1. **The exact hyperboloid function $\eta(x_1, x_2) = x_1 x_2$.**

5. Numerical Experience

The effectiveness of our proposed formulation has been illustrated on three approximation tasks [MW05] and three classification tasks [MW06]. We describe these experiments and their results here. It is important to point out that the present formulation is very different in nature from that presented in [FMS03a] and [MSW04]. Our primary concern here is to incorporate prior knowledge in an explicit and transparent manner without having to kernelize it as was done in [FMS03a] and [MSW04]. In particular, we are able to directly incorporate general implications involving nonlinear inequalities in a linear program by utilizing Theorem 2.1. Synthetic examples were used to show how our approach uses nonlinear prior knowledge to obtain approximations or classifiers that are much better than those obtained without prior knowledge. Although the given prior knowledge for the synthetic examples is strong, the examples illustrate the simplicity and effectiveness of our approach to incorporate prior knowledge into nonlinear support vector classification and approximation. The Wisconsin Prognostic Breast Cancer (WPBC) dataset was used to demonstrate situations in which prior knowledge and data are combined to obtain a better approximation or classifier than by using only prior knowledge or data alone.

5.1. Approximation Datasets. The effectiveness of our proposed approximation formulation (12) has been illustrated on two synthetic datasets and the Wisconsin Prognostic Breast Cancer (WPBC) database, available from [MA92].

5.1.1. *Two-Dimensional Hyperboloid Function.* The first approximation example is the two-dimensional hyperboloid function:

$$\eta(x_1, x_2) = x_1 x_2. \tag{17}$$

This function was studied in [MSW04]. The given data consists of eleven points along the line $x_1 = x_2$, $x_1 \in \{-5, -4, \ldots, 4, 5\}$. The given values at these points are the actual function values.

Figure 1 depicts the two-dimensional hyperboloid function of (17). Figure 2 depicts the approximation of the hyperboloid function by a surface based on the eleven points described above *without* prior knowledge. Figures 1 and 2 are taken from [MSW04].

Figure 3 depicts a much better approximation than that of Figure 2 of the hyperboloid function by a nonlinear surface based on the same eleven points above

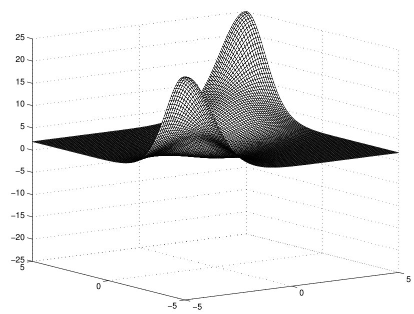

FIGURE 2. **Approximation of the hyperboloid function $\eta(x_1, x_2) = x_1 x_2$ based on eleven exact function values along the line $x_2 = x_1, x_1 \in \{-5, -4, \ldots, 4, 5\}$, but *without* prior knowledge.**

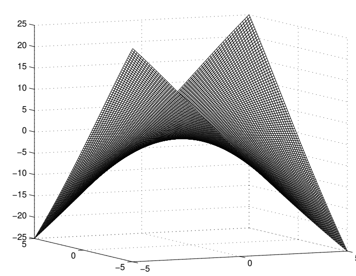

FIGURE 3. **Approximation of the hyperboloid function $\eta(x_1, x_2) = x_1 x_2$ based on the same eleven function values as Figure 2 *plus* prior knowledge consisting of the implication** (18)**.**

plus prior knowledge. The prior knowledge consisted of the implication:

$$x_1 x_2 \leq 1 \Longrightarrow f(x_1, x_2) \leq x_1 x_2, \tag{18}$$

which, because of the nonlinearity of $x_1 x_2$, cannot be handled by [MSW04]. Note that even though the prior knowledge implication (18) provides only partial information regarding the hyperboloid (17) being approximated, applying it is sufficient to improve our kernel approximation substantially as depicted in Figure 3. The prior knowledge implication (18) was applied in its equivalent inequality (5) form, at discrete points as stated in the inequality constraints of (12). In this example, the knowledge was applied at eleven points along the line $x_1 = -x_2$, $x_1 \in \{-5, -4, \ldots, 4, 5\}$.

It is instructive to compare (18) with the prior knowledge used in [MSW04] to obtain a visually similar improvement. In that work, the following prior knowledge was used:

$$\begin{array}{c}(x_1, x_2) \in \{(x_1, x_2) | -\frac{1}{3}x_1 \leq x_2 \leq -\frac{2}{3}x_1\} \Rightarrow \\ f(x_1, x_2) \leq 10x_1\end{array} \tag{19}$$

and

$$\begin{array}{c}(x_1, x_2) \in \{(x_1, x_2) | -\frac{2}{3}x_1 \leq x_2 \leq -\frac{1}{3}x_1\} \Rightarrow \\ f(x_1, x_2) \leq 10x_2.\end{array} \tag{20}$$

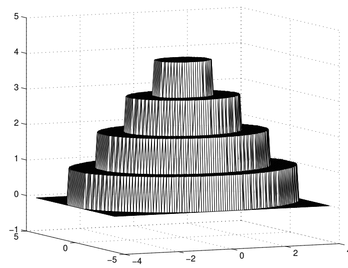

FIGURE 4. **The exact tower function given by** (21).

These implications were implemented by replacing $f(x_1, x_2)$ with its nonlinear kernel approximation (1) and by kernelizing the resulting prior knowledge [MSW04, Equation 18]. The result can then be incorporated into a linear program with no discretization required [MSW04, Proposition 3.1]. However, as is noted in [MSW04], the implications (19) and (20) are not correct everywhere, but are merely intended to coarsely model the global shape of $\eta(x_1, x_2)$. This inexactness arises because of the limitation that knowledge be linear in the input space, and because the use of the nonlinear kernel to map knowledge in the input space to higher dimensions is difficult to interpret. In contrast, the prior knowledge of implication (18) is always correct and exactly captures the shape of the function. Thus, this example illustrates that there is a significant gain in usability due to the fact that the knowledge may be nonlinear in input space features.

5.1.2. *Two-Dimensional Tower Function.* For the second approximation example, we considered the following function:

$$\tau(x_1, x_2) = \begin{cases} 4, \text{ when} & \|(x_1, x_2)\| < 1 \\ 3, \text{ when} & 1 \leq \|(x_1, x_2)\| < 2 \\ 2, \text{ when} & 2 \leq \|(x_1, x_2)\| < 3 \\ 1, \text{ when} & 3 \leq \|(x_1, x_2)\| < 4 \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

which is shown in Figure 4. Due to the visual appearance of this function, we refer to it as the *tower* function.

The data used to approximate the tower function of (21) consists of 400 equally spaced points on the grid $[-4,\ 4] \times [-4,\ 4]$, with given values defined using the following equation:

$$f(x_1, x_2) = \min\{\tau(x_1, x_2), 2\}, \tag{22}$$

where $\tau(x_1, x_2)$ is given by (21). This misleading data explains the chopped-off appearance that is shown by the approximation of Figure 5 which is a poor approximation of the tower function based on this data *without* prior knowledge.

Figure 6 shows an approximation of the tower function using the data described above *plus* the following prior knowledge:

$$(x_1, x_2) \in [-4, 4] \times [-4, 4] \Longrightarrow f(x_1, x_2) = \tau(x_1, x_2), \tag{23}$$

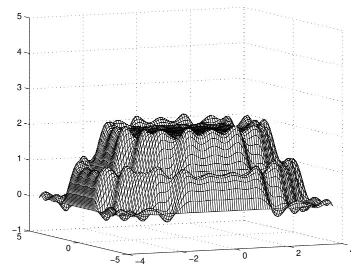

FIGURE 5. **An approximation of the tower function in** (21) **using 400 equally spaced points on** $[-4,4]\times[-4,4]$ **given by** (22) ***without*** **prior knowledge.**

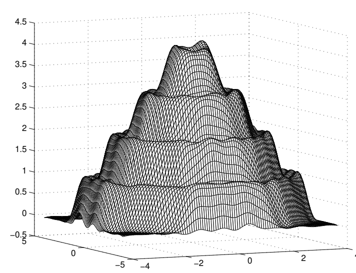

FIGURE 6. **An approximation of the tower function in** (21) **using 400 equally spaced points on** $[-4,4]\times[-4,4]$ **given by** (22) ***with*** **the prior knowledge described in** (23)**.**

where $\tau(x_1, x_2)$ is the exact value of the tower function of (21). This implication was enforced at 2500 equally spaced points on the grid $[-4,4]\times[-4,4]$. The approximation depicted in Figure 6 was made by setting the parameters C and ν of (12) to 10^1 and 10^{20} respectively. Thus, this example illustrates that despite poor initial data, a substantially improved approximation using prior knowledge can be made by incorporating prior knowledge in the form of an implication such as (23).

5.1.3. *Predicting Lymph Node Metastasis.* To demonstrate the effectiveness of our approximation formulation (12) on a real-world dataset, a potentially useful application of knowledge-based approximation to breast cancer prognosis [MSW95, WSHM95, LMW03] was considered. An important prognostic indicator for breast cancer recurrence is the number of metastasized lymph nodes under a patient's armpit which could be as many as 30. To obtain this number, a patient must optionally undergo a potentially debilitating surgery in addition to the removal of the breast tumor. Thus, it is useful to approximate the number of metastasized lymph nodes using available information. The Wisconsin Prognostic Breast Cancer (WPBC) data, in which the primary task is to determine time to recurrence [MA92], contains information on the number of metastasized lymph nodes for 194 breast cancer patients, as well as thirty cytological features obtained by a fine needle aspirate and one histological feature, tumor size, obtained during surgery.

Approximation	RMSE
Without knowledge	5.92
With knowledge	**5.04**
Improvement due to knowledge	14.8%

Table I: Leave-one-out root-mean-squared-error (RMSE) of approximations with and without knowledge on the present WPBC data. Best result is in bold.

Mangasarian et al. demonstrated in [MSW04] that a function that approximated the number of metastasized lymph nodes using four of these features could be improved using prior knowledge. The formulation developed here has been used to approximate the number of metastasized lymph nodes using only the tumor size.

In order to simulate the situation where an expert provides prior knowledge regarding the number of metastasized lymph nodes based on tumor size, the following procedure was used. First, 20% of the data was randomly selected as "past data." This past data was used to develop prior knowledge, while the remaining 80% of the data, the "present data," was used for evaluation. The goal is to simulate the situation in which an expert can provide prior knowledge, but no more data is available. To generate such prior knowledge, kernel approximation was used to find a function $\phi(x) = K(x', B')u - \gamma$, where B is the matrix containing the past data and K is the Gaussian kernel defined in Section 1. This function was then used as the basis for our prior knowledge. Since this function was not believed to be accurate for areas where there was little data in the past data set, this knowledge was imposed only on the region $p(x) \geq 0.1$, where $p(x)$ was the density function for the tumor sizes in B estimated with the `ksdensity` routine, available in the MATLAB statistics toolbox [MAT06]. The following prior knowledge implication was considered:

$$p(x) \geq 0.1 \Longrightarrow f(x) \geq \phi(x) - 0.01. \tag{24}$$

That is, the number of metastasized lymph nodes was greater than the predicted value on the past data, with a tolerance of 0.01. This implication incorporates a typical oncological surgeon's advice that the number of metastasized lymph nodes increases with tumor size. In order to accurately simulate the desired conditions, this knowledge was formed by observing only the past data. No aspect of the prior knowledge was changed after testing began on the present data.

Table I illustrates the improvement resulting from the use of prior knowledge. The first two entries compare the leave-one-out error of function approximations without and with prior knowledge. When training functions on each training set, ten points of the training set were selected as a tuning set. This set was used to choose the value of C from the set $\{2^i | i = -7, \ldots, 7\}$. The kernel parameter was set to 2^{-7}, which gave a smooth curve on the past data set. This value was fixed before testing on the present data. For the approximation *with* knowledge, the parameter ν was set to 10^6, which ensured that the prior knowledge would be taken into account by the approximation. Implication (24) was imposed as prior knowledge, and the discretization for the prior knowledge was 400 equally spaced points on the interval $[1, 5]$. This interval approximately covered the region on which $p(x) \geq 0.1$. We note that the use of prior knowledge led to a 14.8% improvement. In our experience, such an improvement is difficult to obtain in medical tasks, and indicates that the approximation with prior knowledge is more potentially useful than the approximation without prior knowledge.

In order to further illustrate the effectiveness of using prior knowledge, two other experiments were performed. First, the root-mean-squared-error (RMSE) of the function ϕ was calculated on the present data, which was not used to create ϕ. The resulting RMSE was 6.12, which indicates that this function does not, by itself, do a good job predicting the present data. The leave-one-out error on the present data of an approximation that included the present data *and* the past data, but *without* prior knowledge was also calculated. This approach led to less than one percent improvement over the approximation without knowledge shown in Table I, which indicates that the prior knowledge *in the form of the implication (24)* contains more useful information than the *raw* past data alone. These results indicate that the inclusion of the prior knowledge with the present data is responsible for the 14.8% improvement.

5.2. Classification Datasets. The effectiveness of our proposed classification formulation (16) has been illustrated on three publicly available datasets: The Checkerboard dataset [HK96], the Spiral dataset [Wie], and the Wisconsin Prognostic Breast Cancer (WPBC) dataset [MA92].

5.2.1. *Checkerboard Problem.* The first classification example was based on the frequently utilized checkerboard dataset [HK96, Kau99, MM01, LM01, FMS03a]. This synthetic dataset contains two-dimensional points in $[-1, 1] \times [-1, 1]$ labeled so that they form a checkerboard. For this example, a dataset consisting of only the sixteen points at the center of each square in the checkerboard was used to generate a classifier without knowledge. The rows of both matrices A and B of (16) were set equal to the coordinates of the sixteen points, which are the standard values. Figure 7 shows a classifier trained on these sixteen points *without* any additional prior knowledge.

Figure 8 shows a much more accurate classifier trained on the same sixteen points as used in Figure 7, *plus* prior knowledge representing only the leftmost two squares in the bottom row of the checkerboard. This knowledge was imposed via the following implications:

$$\begin{array}{c} -1 \leq x_1 \leq -0.5 \wedge -1 \leq x_2 \leq -0.5 \Longrightarrow f(x_1, x_2) \geq 0, \\ -0.5 \leq x_1 \leq 0 \wedge -1 \leq x_2 \leq -0.5 \Longrightarrow f(x_1, x_2) \leq 0. \end{array} \tag{25}$$

The implication on the first line was imposed at 100 uniformly spaced points in $[-1, -0.5] \times [-1, -0.5]$, and the implication on the second line were imposed at 100 uniformly spaced points in $[-0.5, 0] \times [-1, -0.5]$. No prior knowledge was given for the remaining squares of the checkerboard. We note that this knowledge is very similar to that used in [FMS03a], although our classifier is more accurate here. This example demonstrates that knowledge of the form used in [FMS03a] can be easily applied with our proposed approach without kernelizing the prior knowledge.

5.2.2. *Spiral Problem.* The spiral dataset [Wie, FM01] was used as the second synthetic example. This dataset consists of the two concentric spirals shown in Figure 9.

In order to illustrate the effectiveness of our classification approach (16) on this dataset, labels were provided for only a random subset of the points in Figure 9. For this dataset, the matrix B of Equation (1) consists of all the points in the dataset. Figure 10 shows a classifier trained using ten-fold cross validation on the

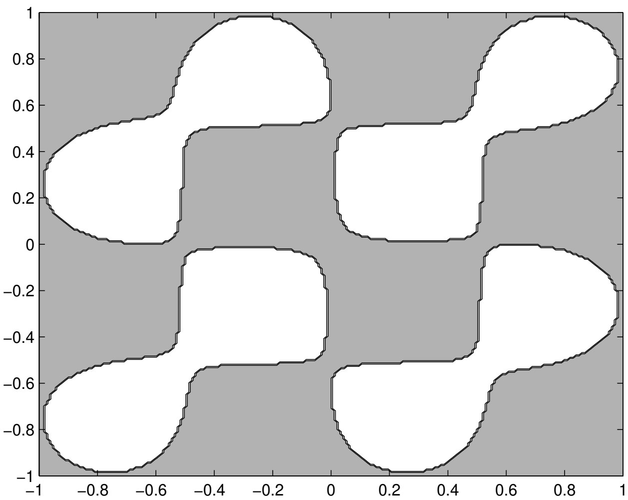

FIGURE 7. **A classifier for the checkerboard dataset trained using only the sixteen points at the center of each square *without* prior knowledge. The white regions denote areas where the classifier returns a value greater than zero, and the gray regions denote areas where the classifier returns a value less than zero. A uniform grid consisting of $40,000$ points was used to create the plot utilizing the obtained classifier.**

points with given labels and no prior knowledge. The points for which labels were given during training are circled. Note that the classifier incorrectly classifies many of the points with label $+1$ for which no label was provided during training.

Figure 11 shows a much more accurate classifier trained on the same labeled points *plus* prior knowledge based on the construction of the spiral dataset. This knowledge can be represented as follows:

(26)
$$\begin{array}{rl} g(x) \leq 0 \Longrightarrow f(x) \geq & 1 \\ g(x) \geq 0 \Longrightarrow f(x) \leq & -1 \quad \text{where} \end{array}$$
$$g(x) = \left\| \begin{pmatrix} \|x\|\cos\left(\frac{\pi(6.5-x)104}{(16)(6.5)}\right) \\ \|x\|\sin\left(\frac{\pi(6.5-x)104}{(16)(6.5)}\right) \end{pmatrix} - x \right\| - \left\| \begin{pmatrix} \|x\|\cos\left(\frac{\pi(6.5-x)104}{(16)(6.5)} + \pi\right) \\ \|x\|\sin\left(\frac{\pi(6.5-x)104}{(16)(6.5)} + \pi\right) \end{pmatrix} - x \right\|.$$

Though complicated in appearance, the derivation of this expression is actually quite straightforward given the source code that generates the spiral dataset [Wie]. To impose the prior knowledge, each implication was imposed at the points defined by the rows of the matrix B for which the left-hand side of the implication held, as well as two additional points near that point. Recall that for this dataset, B contains every point in the dataset as shown in Figure 9. For example, the first implication was imposed on the points x and $x \pm \binom{0.2}{0.2}$ where x is a row of the matrix B and $g(x) \leq 0$.

5.2.3. *Predicting Breast Cancer Survival Time.* We conclude our experimental results with a potentially useful application of the Wisconsin Prognostic Breast

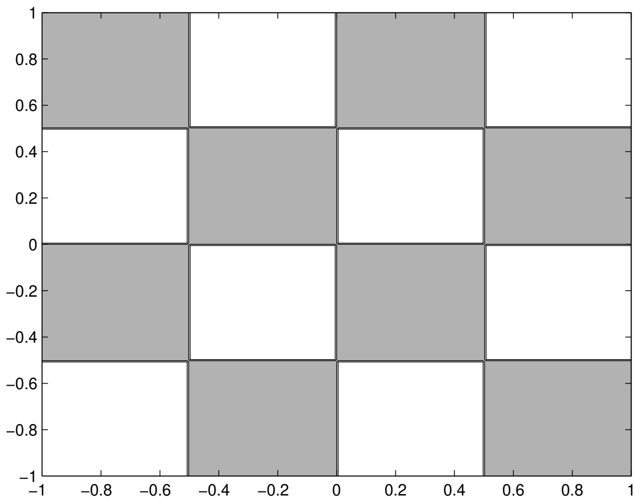

FIGURE 8. **A classifier for the checkerboard dataset trained using the sixteen points at the center of each square *with* prior knowledge representing the two leftmost squares in the bottom row given in (25). The white regions denote areas where the classifier returns a value greater than zero, and the gray regions denote areas where the classifier returns a value less than zero. A uniform grid consisting of $40,000$ points was used to create the plot utilizing the knowledge-based classifier.**

Cancer (WPBC) dataset [MA92, LMW99]. This dataset contains thirty cytological features obtained from a fine needle aspirate and two histological features, tumor size and the number of metastasized lymph nodes, obtained during surgery for breast cancer patients. The dataset also contains the amount of time before each patient experienced a recurrence of the cancer, if any. Here, the task of predicting whether a patient will remain cancer free for at least 24 months is considered. Past experience with this dataset has shown that an accurate classifier for this task is difficult to obtain. In this dataset, 81.9% of patients are cancer free after 24 months. To our knowledge, the best result on this dataset prior to [MW06] is 86.3% correctness obtained by Bennett in [Ben92]. It is possible that incorporating expert information about this task is necessary to obtain higher accuracy on this dataset. In [MW06] it was demonstrated that with sufficient prior knowledge, our approach can achieve 91.0% correctness.

To obtain prior knowledge for this dataset, the number of metastasized lymph nodes was plotted against the tumor size, along with the class label, for each patient. An oncological surgeon's advice was then simulated by selecting regions containing patients who experienced a recurrence withing 24 months. In a typical machine learning task, not all of the class labels would be available. However, the purpose here is to demonstrate that if an expert is able to provide useful prior knowledge, our approach can effectively apply that knowledge to learn a more accurate classifier. We leave studies on this dataset in which an expert provides knowledge without all

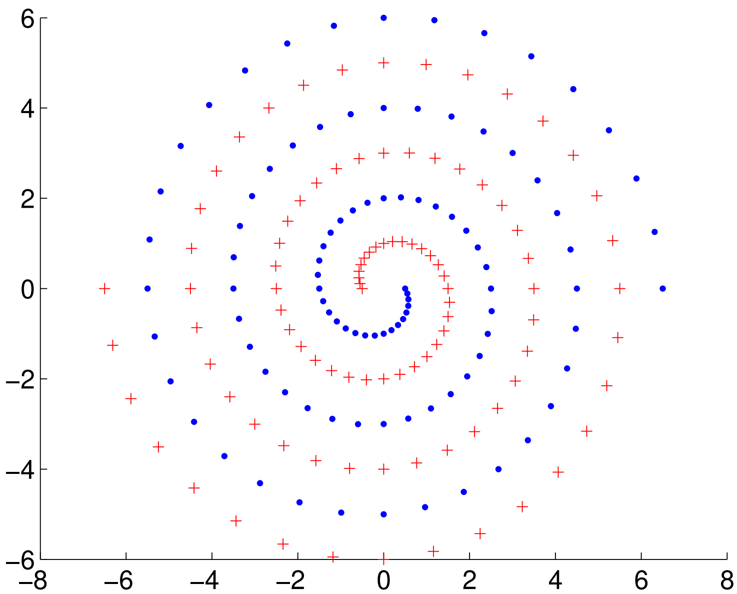

FIGURE 9. **The spiral dataset. The filled circles represent points with label $+1$ while the crosses represent points with label -1.**

of the labels available to future work. In such studies, the expert would be given information regarding the class of only data points in a training set that is a subset of all the data, and then give advice on the class of points in the entire dataset. The prior knowledge constructed for this dataset is depicted in Figure 12 and consists of the following three implications:

(27)
$$
\begin{array}{ll}
\left\| \binom{(5.5)x_1}{x_2} - \binom{(5.5)7}{9} \right\| + \left\| \binom{(5.5)x_1}{x_2} - \binom{(5.5)4.5}{27} \right\| - 23.0509 & \le 0 \Longrightarrow f(x) \ge 1 \\
\begin{pmatrix} -x_2 + 5.7143x_1 - 5.75 \\ x_2 - 2.8571x_1 - 4.25 \\ -x_2 + 6.75 \end{pmatrix} & \le 0 \Longrightarrow f(x) \ge 1 \\
\frac{1}{2}(x_1 - 3.35)^2 + (x_2 - 4)^2 - 1 & \le 0 \Longrightarrow f(x) \ge 1.
\end{array}
$$

The class $+1$ represents patients who experienced a recurrence in less than 24 months. Here, x_1 is the tumor size, and x_2 is the number of metastasized lymph nodes. Each implication is enforced at the points in the dataset for which the left-hand side of the implication is true. These regions are shown in Figure 12. The first implication corresponds to the region closest to the upper right-hand corner. The triangular region corresponds to the second implication, and the small elliptical region closest to the x_1 axis corresponds to the third implication. Although these implications seem complicated, it would not be difficult to construct a more intuitive interface similar to standard graphics programs to allow a user to create arbitrary regions. Applying these regions with our approach would be straightforward.

In order to evaluate our proposed approach, the misclassification rates of two classifiers on this dataset were compared. One classifier is learned without prior

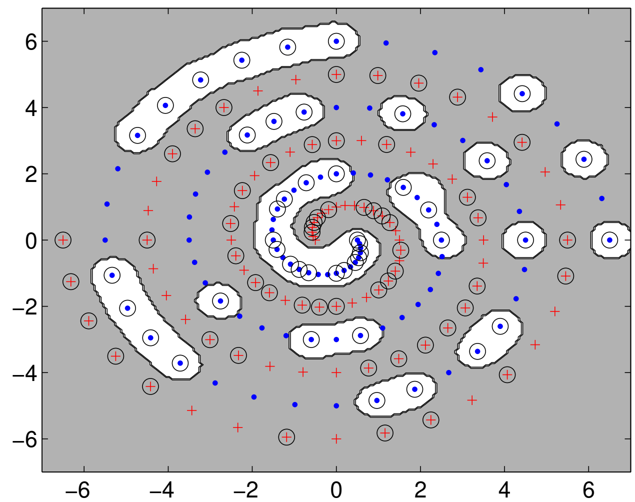

FIGURE 10. **A classifier for the spiral dataset trained using only a subset of given labels *without* prior knowledge. The circled points represent the labeled points constituting the dataset represented by the matrix A of Equation (16). The matrix B consists of all points shown. The white regions denote areas where the classifier returns a value greater than zero, thus classifying the points therein as $+1$, i.e. white $\Rightarrow$ dots. Gray regions denote areas where the classifier returns a value less than zero, thus classifying points as -1, i.e. gray $\Rightarrow$ crosses. Note the many points (dots) incorrectly classified as -1.**

knowledge, while the second classifier is learned using the prior knowledge given in (27). For both cases the rows of the matrices A and B of (16) were set to the usual values, that is to the coordinates of the points of the training set. The misclassification rates are computed using leave-one-out cross validation. For each fold, the parameter ν and the kernel parameter μ were chosen from the set $\{2^i | i \in \{-7, \ldots, 7\}\}$ by using ten-fold cross validation on the training set of the fold. In the classifier with prior knowledge, the parameter σ was set to 10^6, which corresponds to very strict adherence to the prior knowledge. The results are summarized in Table 1. The reduction in misclassification rate indicates that our classification approach can use appropriate prior knowledge to obtain a classifier on this difficult dataset with 50% improvement.

6. Conclusion and Outlook

We have proposed a unified, computationally effective framework for handling general nonlinear prior knowledge in kernel approximation and kernel classification problems. We have reduced such prior knowledge to easily implemented linear constraints in a linear programming formulation. We have demonstrated the effectiveness of our approach on four synthetic problems and two important real world problems arising in breast cancer prognosis. Possible future extensions are to even

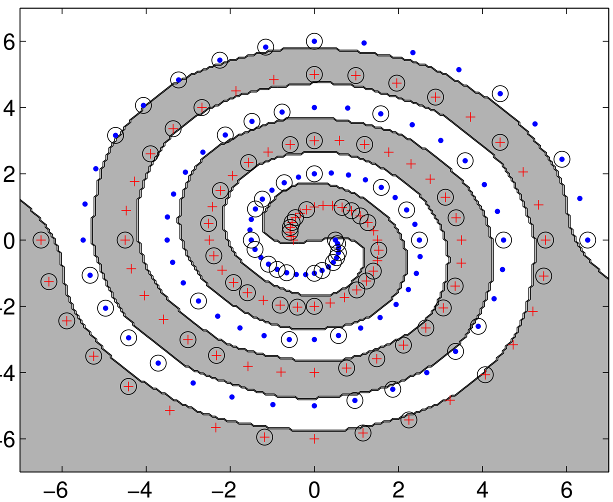

FIGURE 11. **A classifier for the spiral dataset trained using only the subset represented by circled points with given labels *plus* the prior knowledge given in (26). The white regions denote areas where the classifier returns a value greater than zero and should contain only dots. The gray regions denote areas where the classifier returns a value less than zero and should contain only crosses. Note that there are no misclassified points.**

Classifier	Misclassification Rate
Without knowledge	0.1806
With knowledge	**0.0903**
Improvement due to knowledge	50.0%

TABLE 1. **Leave-one-out misclassification rate of classifiers with and without knowledge on the WPBC (24 mo.) dataset. Best result is in bold.**

more general prior knowledge, such as that where the right hand side of the implications (3) and (7) are replaced by very general nonlinear inequalities involving the kernel function (1). Another important avenue of future work is to construct an interface which allows users to easily specify arbitrary regions to be used as prior knowledge.

Acknowledgments. The research described in this Data Mining Institute Report 06-06, November 2006, was supported by National Science Foundation Grants CCR-0138308 and IIS-0511905, the Microsoft Corporation and ExxonMobil.

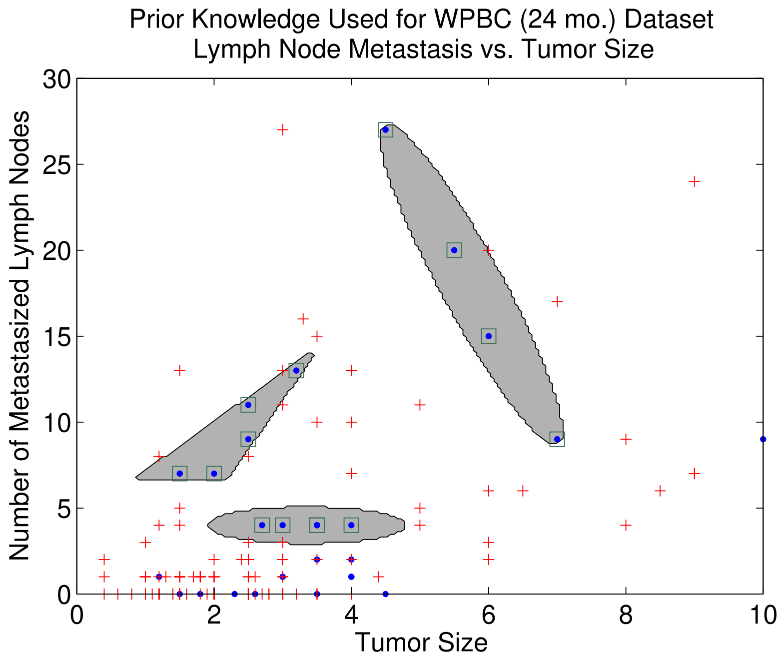

FIGURE 12. **Number of metastasized lymph nodes versus tumor size for the WPBC (24 mo.) dataset. The solid dots represent patients who experienced a recurrence within 24 months of surgery, while the crosses represent the cancer free patients. The shaded regions which correspond to the areas in which the left-hand side of one of the three implications in Equation (27) is true simulate an oncological surgeon's prior knowledge regarding patients that are likely to have a recurrence. Prior knowledge was enforced at the points enclosed in squares.**

References

[Ben92] K. P. Bennett, *Decision tree construction via linear programming*, Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference (Utica, Illinois) (M. Evans, ed.), 1992, pp. 97–101.

[BM98] P. S. Bradley and O. L. Mangasarian, *Feature selection via concave minimization and support vector machines*, Machine Learning Proceedings of the Fifteenth International Conference(ICML '98) (San Francisco, California) (J. Shavlik, ed.), Morgan Kaufmann, 1998, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps, pp. 82–90.

[CM98] V. Cherkassky and F. Mulier, *Learning from data - concepts, theory and methods*, John Wiley & Sons, New York, 1998.

[CST00] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines*, Cambridge University Press, Cambridge, 2000.

[FM01] G. Fung and O. L. Mangasarian, *Proximal support vector machine classifiers*, Proceedings KDD-2001: Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA (New York) (F. Provost and R. Srikant, eds.), Association for Computing Machinery, 2001, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps, pp. 77–86.

[FMS02] G. Fung, O. L. Mangasarian, and A. Smola, *Minimal kernel classifiers*, Journal of Machine Learning Research (2002), 303–321, University of Wisconsin Data Mining Institute Technical Report 00-08, November 200, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-08.ps.

[FMS03a] G. Fung, O. L. Mangasarian, and J. Shavlik, *Knowledge-based nonlinear kernel classifiers*, Tech. 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/03-02.ps. *Conference on Learning Theory (COLT 03) and Workshop on Kernel Machines*, Washington D.C., August 24-27, 2003. Proceedings edited by M. Warmuth and B. Schölkopf, Springer Verlag, Berlin, 2003, 102-113.

[FMS03b] ______, *Knowledge-based support vector machine classifiers*, Advances in Neural Information Processing Systems 15 (Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, eds.), MIT Press, Cambridge, MA, October 2003, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-09.ps, pp. 521–528.

[GL98] M. A. Goberna and M. A. López, *Linear semi-infinite optimization*, John Wiley, New York, 1998.

[HK96] T. K. Ho and E. M. Kleinberg, *Checkerboard dataset*, 1996, http://www.cs.wisc.edu/math-prog/mpml.html.

[HL04] S.Y. Huang and Y.-J. Lee, *Theoretical study on reduced support vector machines*, Technical report, National Taiwan University of Science and Technology, Taipei, Taiwan, 2004, yuh-jye@mail.ntust.edu.tw.

[Kau99] L. Kaufman, *Solving the quadratic programming problem arising in support vector classification*, Advances in Kernel Methods - Support Vector Learning (B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds.), MIT Press, 1999, pp. 147–167.

[LM01] Y.-J. Lee and O. L. Mangasarian, *RSVM: Reduced support vector machines*, Proceedings of the First SIAM International Conference on Data Mining, Chicago, April 5-7, 2001, CD-ROM, 2001, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-07.ps.

[LMW99] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, *Breast cancer survival and chemotherapy: a support vector machine analysis*, Tech. Report 99-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, December 1999, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Volume 55, 2000, 1-10. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-10.ps.

[LMW03] ______, *Survival-time classification of breast cancer patients*, Computational Optimization and Applications **25** (2003), 151–166, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-03.ps.

[LSG06] Q. V. Le, A. J. Smola, and T. Gärtner, *Simpler knowledge-based support vector machines*, Proceedings of the 23^{rd} International Conference on Machine Learning, Pittsburgh, PA, 2006, 2006, http://www.icml2006.org/icml2006/technical/accepted.html.

[MA92] P. M. Murphy and D. W. Aha, *UCI machine learning repository*, 1992, www.ics.uci.edu/∼mlearn/MLRepository.html.

[Man69] O. L. Mangasarian, *Nonlinear programming*, McGraw–Hill, New York, 1969, Reprint: SIAM Classic in Applied Mathematics 10, 1994, Philadelphia.

[Man00] ______, *Generalized support vector machines*, Advances in Large Margin Classifiers (Cambridge, MA) (A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, eds.), MIT Press, 2000, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps, pp. 135–146.

[MAT06] MATLAB, *User's guide*, The MathWorks, Inc., Natick, MA 01760, 1994-2006, http://www.mathworks.com.

[MM01] O. L. Mangasarian and D. R. Musicant, *Lagrangian support vector machines*, Journal of Machine Learning Research **1** (2001), 161–177, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-06.ps.

[MST+05] R. Maclin, J. Shavlik, L. Torrey, T. Walker, and E. Wild, *Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression*, Proceedings of the 20th National Conference on Artificial Intelligence, 2005, pp. 819–824.

[MSW95] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, *Breast cancer diagnosis and prognosis via linear programming*, Operations Research **43** (1995), no. 4, 570–577.

[MSW04] O. L. Mangasarian, J. W. Shavlik, and E. W. Wild, *Knowledge-based kernel approximation*, Journal of Machine Learning Research **5** (2004), 1127–1141, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/03-05.ps.

[MSWT06] R. Maclin, J. Shavlik, T. Walker, and L. Torrey, *A simple and effective method for incorporating advice into kernel methods*, Proceedings of the 21st National Conference on Artificial In telligence, 2006.

[MW05] O. L. Mangasarian and E. W. Wild, *Nonlinear knowledge in kernel approximation*, Tech. Report 05-05, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, October 2005, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/05-05.pdf. IEEE Transactions on Neural Networks, to appear.

[MW06] ———, *Nonlinear knowledge-based classification*, Tech. Report 06-04, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, August 2006, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/06-04.pdf.

[SS02] B. Schölkopf and A. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[Vap00] V. N. Vapnik, *The nature of statistical learning theory*, second ed., Springer, New York, 2000.

[Wie] A. Wieland, *Twin spiral dataset*, http://www-cgi.cs.cmu.edu/afs/cs.cmu.edu/project/ai-repository/ai/areas/neural/bench/cmu/0.html.

[WSHM95] W. H. Wolberg, W. N. Street, D. N. Heisey, and O. L. Mangasarian, *Computerized breast cancer diagnosis and prognosis from fine-needle aspirates*, Archives of Surgery **130** (1995), 511–516.

[ZRHT04] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, 1-*Norm support vector machines*, Advances in Neural Information Processing Systems 16–NIPS2003 (Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, eds.), MIT Press, 2004.

Computer Sciences Department, University of Wisconsin, Madison, WI 53706 and Department of Mathematics, University of California at San Diego, La Jolla, CA 92093.
E-mail address: `olvi@cs.wisc.edu`

Computer Sciences Department, University of Wisconsin, Madison, WI 53706.
E-mail address: `wildt@cs.wisc.edu`