Nonlinear Knowledge in Kernel Approximation

O. L. Mangasarian & E. W. Wild

***Abstract*— Prior knowledge over arbitrary general sets is incorporated into nonlinear kernel approximation problems in the form of 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 formulation is demonstrated on two synthetic examples and an important lymph node metastasis prediction problem. All these problems exhibit marked improvements upon the introduction of prior knowledge over nonlinear kernel approximation approaches that do not utilize such knowledge.**

I. INTRODUCTION

Prior knowledge has been used effectively in improving classification both for linear [1] and nonlinear [2] kernel classifiers as well as for nonlinear kernel approximation [3], [4]. 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, in the present work no kernelization of the prior knowledge sets is used in order to incorporate that knowledge into a nonlinear classifier or a function approximation. Furthermore, the region in the input space on which the prior knowledge is given is completely arbitrary in the present work, whereas in all previous work it had to be given on convex polyhedral sets. The present approach is possible through the use of a fundamental theorem of the alternative for convex functions that we describe in Section II of the paper, whereas previous work utilized such a theorem for linear inequalities *only*.

In Section III we describe our linear programming formulation that incorporates nonlinear prior knowledge into a nonlinear kernel, while Section IV gives numerical examples that show prior knowledge can improve a nonlinear kernel approximation significantly. Section V concludes the paper.

We describe our notation now. All vectors will be column vectors unless transposed to a row vector by the prime notation $'$. 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: $(\sum_{i=1}^{n}(x_i)^2)^{\frac{1}{2}}$. 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, $K(x', A')$ is a row vector in R^m and $K(A, A')$ is an $m \times m$ matrix. We shall make no assumptions on our kernels other than symmetry, that is $K(A, B)' = K(B', A')$, and in particular we shall not assume or make use of Mercer's positive definiteness condition [5], [6], [7], [8]. The base of the natural logarithm will be denoted by ε. A frequently used kernel in nonlinear classification is the Gaussian kernel [5], [9], [8] 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. Approximate equality is denoted by $\approx$, while the abbreviation "s.t." stands for "subject to".

II. CONVERSION OF NONLINEAR PRIOR KNOWLEDGE INTO LINEAR CONSTRAINTS

The problem that we wish to impart prior knowledge to consists of approximating a function f from R^n to R for which approximate or exact function values are given on a dataset of m points in R^n denoted by rows of the matrix $A \in R^{m\times n}$. Thus, corresponding to each point A_i we are given an exact or inexact value of f, denoted by a real number y_i, $i = 1, \ldots, m$. We wish to approximate f by a nonlinear kernel function as follows:

$$K(x', A')\alpha + b, \tag{1}$$

where, $K(x', A') : R^{1\times n} \times R^{n\times m} \longrightarrow R^{1\times m}$ is an arbitrary kernel and $\alpha \in R^m$ and $b \in R$ are parameters to be determined such that:

$$K(A, A')\alpha + be - y \approx 0, \tag{2}$$

and such that some prior knowledge is also utilized in the construction of our approximation $K(x', A')\alpha + b$ through the following implication:

$$g(x) \leq 0 \implies K(x', A')\alpha + b \geq h(x). \tag{3}$$

Here, $g(x) : \Gamma \subset R^n \longrightarrow R^k$ is a k-dimensional function defined on a subset Γ of R^n that determines the region in the input space where prior knowledge requires that the

The research described in this Data Mining Institute Report 05-05, was supported by National Science Foundation Grants CCR-0138308 and IIS-0511905.

O. L. Mangasarian & E. W. Wild are with the Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706. E-mail: olvi@cs.wisc.edu, wildt@cs.wisc.edu.

approximating function $K(x', A')\alpha + b$ be larger than some known function $h(x) : \Gamma \subset R^n \longrightarrow R$. In previous work [2], [3] prior knowledge implications such as (3) could not be handled as we shall do here by using Theorem 2.1 below. Instead, in [2], [3], 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, all previous work [2], [3] could handle only linear $g(x)$ and $h(x)$. The implication (3) can be written in the following equivalent logical form:

$$\begin{array}{c} g(x) \leq 0,\ K(x', A')\alpha + b - h(x) < 0, \\ \text{has no solution } x \in \Gamma. \end{array} \tag{4}$$

It is precisely implication (3) that we shall try to convert to a system of inequalities which is linear in the approximating function parameters (α, b) 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 not both.

Theorem 2.1: **Prior Knowledge as System of Linear Inequalities** For a fixed $\alpha \in R^n, b \in R$, the following are equivalent:

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

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

$$K(x', A')\alpha + b - h(x) + v'g(x) \geq 0,\ \forall x \in \Gamma, \tag{5}$$

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

Proof (i)$\Longrightarrow$(ii): This follows from [10, Corollary 4.2.2], the fact that the functions $g(x)$ and $K(x', A')\alpha + b - h(x)$ 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 $g(x) \leq 0,\ K(x', A')\alpha + b - h(x) < 0$, which would result in the contradiction:

$$0 > K(x', A')\alpha + b - h(x) + v'g(x) \geq 0.\square \tag{6}$$

We note immediately that in the proposed application in Section III of converting prior knowledge to linear inequalities in the parameters (α, b) all we need is the implication (i)$\Longleftarrow$(ii), which **requires no assumptions whatsoever** on the functions $g(x)$, $K(x', A')$, $h(x)$ or on the parameter α. *However, we also note that even though we do not make explicit use of the necessity part of Theorem 2.1, it is quite important to have such a result in order to show that* (5) *is not a vacuous sufficient condition since it does indeed hold under certain assumptions.*

We further note that the implication (3) can represent fairly complex knowledge such as $K(x', A')\alpha + b$ being equal to any desired function whenever $g(x) \leq 0$.

We turn now to our linear programming formulation of the knowledge-based nonlinear kernel approximation.

III. NONLINEAR PRIOR KNOWLEDGE APPROXIMATION VIA LINEAR PROGRAMMING

We first formulate the approximation (2) without knowledge as follows. We measure the error in (2) by a vector $s \in R^m$ defined by:

$$-s \leq K(A, A')\alpha + be - y \leq s. \tag{7}$$

We now drive this error down by minimizing the 1-norm of the error s together with the 1-norm of α for complexity reduction or stabilization. This leads to the following constrained optimization problem with positive parameter C that determines the relative weight of exact data fitting to complexity reduction:

$$\begin{array}{rl} \displaystyle\min_{(\alpha,b,s)} & \|\alpha\|_1 + C\|s\|_1 \\ \text{s.t.} & -s \leq K(A, A')\alpha + be - y \leq s, \end{array} \tag{8}$$

which can be represented as the following linear program:

$$\begin{array}{rrcccl} \displaystyle\min_{(\alpha,b,s,a)} & & & e'a + Ce's & & \\ \text{s.t.} & -s & \leq & K(A, A')\alpha + be - y & \leq & s, \\ & -a & \leq & \alpha & \leq & a. \end{array} \tag{9}$$

We now introduce prior knowledge contained in the implication (3) by making use of Theorem 2.1 and converting it to the linear constraints (5) and incorporating into the linear program (9) as follows:

$$\begin{array}{rl} \displaystyle\min_{(\alpha,b,s,a,v,z_1,\ldots,z_\ell)} & e'a + Ce's + \nu\displaystyle\sum_{i=1}^{\ell} z_i \\ \text{s.t.} & -s \leq K(A, A')\alpha + be - y \leq s, \\ & -a \leq \alpha \leq a, \\ & K(x^{i'}, A')\alpha + b - h(x^i) + v'g(x^i) + z_i \geq 0, \\ & v \geq 0,\ z_i \geq 0,\ i = 1, \ldots, \ell. \end{array} \tag{10}$$

We note that we have discretized the variable $x \in \Gamma$ in the next to the last constraint above to a mesh of points $x^1, x^2, \ldots, x^\ell$ in order to convert a semi-infinite linear program [11], that is a linear program with an infinite number of constraints, to a linear program with a finite number of constraints. We have also added nonnegative slack error variables z_i, $i = 1, \ldots, \ell$, to allow small deviations in satisfying the prior knowledge. The sum of these nonnegative slack variables $z_1, z_2, \ldots, z_\ell$ for the prior knowledge inequalities are minimized with weight $\nu > 0$ in the objective function in order to drive them to zero to the extent possible. Thus, the magnitude of the parameter ν enforces prior knowledge while the magnitude of C enforces data fitting. Note that ν and C can be thought of as hyperparameters in a Bayesian setting.

We turn now to computational results and test examples of the proposed approach for incorporating nonlinear knowledge into kernel approximation problems.

IV. COMPUTATIONAL RESULTS

To illustrate the effectiveness of our proposed formulation, we report results on two synthetic datasets and the Wisconsin Prognostic Breast Cancer (WPBC) database, available from [12]. It is important to point out that the present formulation is very different in nature from that presented by Mangasarian et al. in [3]. Our concern here is to demonstrate uses of more complex prior knowledge that could not be handled exactly in [3]. In particular, we are able to incorporate general implications involving nonlinear functions as linear inequalities in

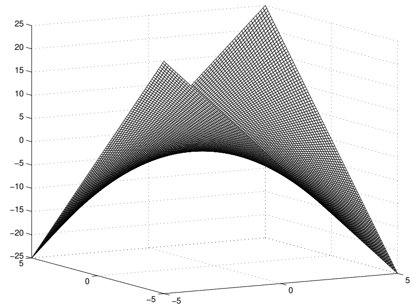

Fig. 1. **The exact hyperboloid function $f(x_1, x_2) = x_1 x_2$.**

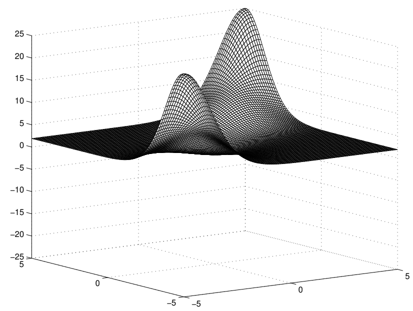

Fig. 2. **Approximation of the hyperboloid function $f(x_1, x_2) = x_1 x_2$ based on eleven exact function values along the line $x_2 = x_1, x_1 \in \{-5, -4, \ldots, 4, 5\}$, but *without* prior knowledge.**

a linear program by utilizing Theorem 2.1. Our two synthetic examples will show how our approach uses nonlinear prior knowledge to obtain approximations that are much better than those obtained without prior knowledge. The prior knowledge may be strong, but the examples demonstrate that it can be easily and correctly incorporated into our formulation to improve the obtained approximation. We will also use the WPBC dataset to demonstrate a situation where prior knowledge and conventional data are combined to obtain a better approximation than that by using only prior knowledge or data alone.

A. TWO-DIMENSIONAL HYPERBOLOID FUNCTION

Our first example is the two-dimensional hyperboloid function:

$$f(x_1, x_2) = x_1 x_2. \tag{11}$$

This function was studied in [3]. 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 (11). 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 [3].

Figure 3 depicts a much better approximation of the hyperboloid function by a nonlinear surface based on the same eleven points above *plus* prior knowledge. The prior knowledge consisted of the implication:

$$x_1 x_2 \leq 1 \Longrightarrow f(x_1, x_2) \leq x_1 x_2, \tag{12}$$

which, because of the nonlinearity of $x_1 x_2$, cannot be handled by [3] nor by any other approximation procedure that we are

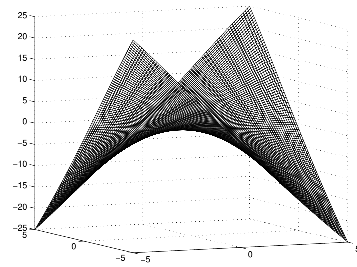

Fig. 3. **Approximation of the hyperboloid function $f(x_1, x_2) = x_1 x_2$ based on the same eleven function values as Figure 2 *plus* prior knowledge consisting of the implication (12).**

aware of. Note that even though the prior knowledge implication (12) provides only partial information regarding the hyperboloid (11) being approximated, applying it is sufficient to improve our kernel approximation substantially as depicted in Figure 3. We applied this prior knowledge implication (12), in its equivalent inequality (5) form, at discrete points as stated in the last three inequality constraints of (10). 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 (12) with the prior knowledge used in [3] to obtain a visually similar improvement. In that work, the following prior knowledge was used:

$$\begin{array}{c} (x_1, x_2) \in \{(x_1, x_2) | -\frac{1}{3}x_1 \leq x_2 \leq -\frac{2}{3}x_1\} \Rightarrow \\ f(x_1, x_2) \leq 10x_1 \end{array} \tag{13}$$

and

$$\begin{array}{c} (x_1, x_2) \in \{(x_1, x_2) | -\frac{2}{3}x_1 \leq x_2 \leq -\frac{1}{3}x_1\} \Rightarrow \\ f(x_1, x_2) \leq 10x_2. \end{array} \tag{14}$$

These implications were implemented by replacing $f(x_1, x_2)$ with its nonlinear kernel approximation (1) and by kernelizing the resulting prior knowledge [3, Equation 18]. The result can then be incorporated into a linear program with no discretization required [3, Proposition 3.1]. However, as is noted in [3], the implications (13) and (14) are not correct everywhere, but are merely intended to coarsely model the global shape of $f(x_1, x_2)$. 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. Here, in contrast, the prior knowledge of implication (12) 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.

B. TWO-DIMENSIONAL TOWER FUNCTION

For our second example, we consider the following function:

$$g(x_1, x_2) = \begin{cases} 4, \text{ when} & \|(x_1, x_2)\| < 1 \\ 3, \text{ when} & 1 \leq \|(x_1, x_2)\| < 2 \\ 2, \text{ when} & 2 \leq \|(x_1, x_2)\| < 3 \\ 1, \text{ when} & 3 \leq \|(x_1, x_2)\| < 4 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

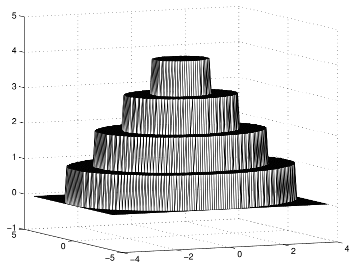

Fig. 4. **The exact tower function given by** (15)**.**

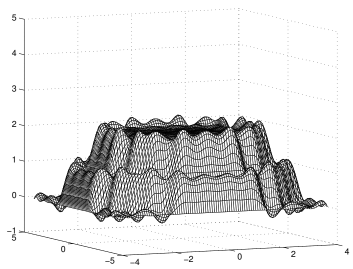

Fig. 5. **An approximation of the tower function in** (15) **using 400 equally spaced points on** $[-4,4]\times[-4,4]$ **given by** (16) ***without*** **prior knowledge.**

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 (15) consists of 400 equally spaced points on the grid $[-4,\ 4]\ \times\ [-4,\ 4]$, with given values defined using the following equation:

$$f(x_1, x_2) = \min\{g(x_1, x_2), 2\}, \tag{16}$$

where $g(x_1, x_2)$ is given by (15). 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:

$$(x_1, x_2) \in [-4,4]\times[-4,4] \Longrightarrow f(x_1, x_2) = g(x_1, x_2), \tag{17}$$

where $g(x_1, x_2)$ is the exact value of the tower function of (15). 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 (10) 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 (17). Such incorporation of prior knowledge, involving nonlinear functions, as linear constraints in an optimization problem has not been made previously.

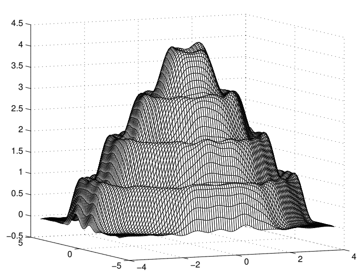

Fig. 6. **An approximation of the tower function in** (15) **using 400 equally spaced points on** $[-4,4]\times[-4,4]$ **given by** (16) ***with*** **the prior knowledge described in** (17)**.**

C. PREDICTING LYMPH NODE METASTASIS

To conclude our numerical results, we consider a potentially useful application of knowledge-based approximation to breast cancer prognosis [13], [14], [15]. 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 [12], 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. Mangasarian et al. demonstrated in [3] that a function that approximated the number of metastasized lymph nodes using four of these features could be improved using prior knowledge. We shall use the formulation developed here 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, we used the following procedure. First, we randomly selected 20% of the data 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, we used kernel approximation to find a function $f_1(x) = K(x', A^{1'})\alpha^1 + b^1$, where A^1 is the matrix containing the past data and K is the Gaussian kernel defined in Section I. We then used this function as the basis for our prior knowledge. Since we did not believe that this function was accurate for areas where there was little data in the past data set, we imposed this knowledge only on the region $p(x) \geq 0.1$, where $p(x)$ was the density function for the tumor sizes in A^1 estimated with the `ksdensity` routine, available in the MATLAB statistics toolbox [16]. We considered the following prior knowledge implication:

$$p(x) \geq 0.1 \Longrightarrow f(x) \geq f_1(x) - 0.01. \tag{18}$$

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

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.

surgeon's advice that the number of metastasized lymph nodes increases with tumor size. In order to accurately simulate the desired conditions, we formed this knowledge by observing only the past data. We did not change any aspect of the prior knowledge after we began testing 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 we observed 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 (18) 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, we also performed two other experiments. First, we calculated the root-mean-squared-error (RMSE) of the function f_1 on the present data, which was not used to create f_1. The resulting RMSE was 6.12, which indicates that this function does not, by itself, do a good job predicting the present data. We also calculated 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. 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 (18)* 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.

V. CONCLUSION AND OUTLOOK

We have proposed a computationally effective framework for handling general nonlinear prior knowledge in kernel approximation 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 two synthetic problems and an important real world problem arising in breast cancer prognosis. Possible future extensions are to even more general prior knowledge, such as that where the right hand side of the implication (3) is replaced by a very general nonlinear inequality. Another fruitful avenue of research would be to apply the general nonlinear kernel knowledge to classification problems, for which prior approaches involved unnecessary kernelization of the prior knowledge.

REFERENCES

[1] G. Fung, O. L. Mangasarian, and J. Shavlik, "Knowledge-based support vector machine classifiers," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, Cambridge, MA, October 2003, pp. 521–528, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-09.ps.

[2] ——, "Knowledge-based nonlinear kernel classifiers," in *Conference on Learning Theory (COLT 03) and Workshop on Kernel Machines*, M. Warmuth and B. Schölkopf, Eds. Berlin: Springer–Verlag, 2003, pp. 102–113, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/03-02.ps.

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

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

[5] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer, 2000.

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

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

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

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

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

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

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

[13] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570–577, July-August 1995.

[14] 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*, vol. 130, pp. 511–516, 1995.

[15] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, "Survival-time classification of breast cancer patients," *Computational Optimization and Applications*, vol. 25, pp. 151–166, 2003, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-03.ps.

[16] MATLAB, *User's Guide*. Natick, MA 01760: The MathWorks, Inc., 1994-2001, http://www.mathworks.com.