Abstract:

In this article we study support vector machine (SVM) classifiers in the face of uncertain knowledge sets and show how data uncertainty in knowledge sets can be treated in SVM classification by employing robust optimization. We present knowledge-based SVM classifiers with uncertain knowledge sets using convex quadratic optimization duality. We show that the knowledge-based SVM, where prior knowledge is in the form of uncertain linear constraints, results in an uncertain convex optimization problem with a set containment constraint. Using a new extension of Farkas' lemma, we reformulate the robust counterpart of the uncertain convex optimization problem in the case of interval uncertainty as a convex quadratic optimization problem. We then reformulate the resulting convex optimization problems as a simple quadratic optimization problem with non-negativity constraints using the Lagrange duality. We obtain the solution of the converted problem by a fixed point iterative algorithm and establish the convergence of the algorithm. We finally present some preliminary results of our computational experiments of the method

Keywords: robust optimization | robust Farkas' lemma | support vector machines | uncertain knowledge sets | quadratic optimization | duality | 65K10 | 90C25 | 90M45

Article:

1. Introduction

Support vector machines (SVMs) [3,20] are an optimization-based solution method for data classification problems. The SVM models are generally formulated as linear or convex quadratic programming problems. The knowledge-based SVM formulation generates separating hyperplanes by training on data and utilizing prior knowledge [16,19]. Incorporating prior knowledge into SVMs in the form of knowledge sets often improves correctness of the classifier or reduce the amount of training data needed. Knowledge-based SVM approaches have been successfully examined in many recent studies [12–14,16], where knowledge sets are assumed to be known with certainty. In reality, however, they are inherently performed under uncertainty because the data inputs of prior expert knowledge, such as doctor's experience, often suffer from experimental or prediction errors. Consequently, it is of great interest to examine the ways of developing SVM classifiers that are capable of handling data uncertainty in knowledge-based classification and mining.

In this article, we study knowledge-based SVMs within the framework of robust optimization that incorporates prior knowledge in the form of uncertain linear constraints. Robust optimization [1] has emerged as a powerful approach for dealing with data uncertainty and it treats uncertainty as deterministic, but does not limit data values to point estimates. In this framework, one associates with the uncertain optimization problem its *robust counterpart* [2,9,11,15] where the uncertain constraints are enforced for every possible value of the data within their prescribed uncertainty sets.

Key to our approach is the reformulation of the robust counterpart of an uncertain knowledge-based SVM model as a convex quadratic optimization problem using a generalized Farkas' lemma. The reformulated problem is then simplified as a quadratic optimization problem with non-negativity constraints using the Lagrange duality. A solution of the simplified quadratic problem is then obtained by a fixed-point iterative algorithm. Our approach extends the method of simultaneous classification and feature selection of [5,21,22], which recently led to the development of a screening algorithm for HIV-associated neurocognitive disorders[4].

The outline of this article is a follows. Section 2 presents preliminaries on knowledge-based SVMs. Section 3 develops a generalization of the Farkas lemma to systems of uncertain linear inequalities. Section 4 formulates the robust knowledge-based SVM as a convex quadratic programming problem. Section 5 describes duality and converts the robust SVM as a simple quadratic optimization problem with non-negativity constraints. It also presents a fixed-point pseudo-algorithm and its convergence to the solution of the robust SVM. Section 6 gives preliminary results on the computational experiments of the method. Section 7 concludes with a discussion on further research.

2. Preliminaries on knowledge-based SVMs

The conventional SVM problem is formulated as discriminating between m data points in $\mathbb{R}^n$. The points are stored in an $m \times n$matrix A, with the ith point a_i stored on the ith row of A. Each point is defined to be belonging to either class $\mathcal{A}$ or $\mathcal{B}$, which is recorded along the diagonal of the diagonal matrix $D \in \mathbb{R}^{m\times m}$. The diagonal elements $D_{ii} = +1$, if the point a_i belongs to $\mathcal{A}$, and $D_{ii} = -1$, if the point belongs to $\mathcal{B}$.

We discriminate between the two data sets with the hyperplane:

$$\{a \in \mathbb{R}^n : a^T w = \gamma\} \qquad (1)$$

Naturally, if the convex hulls of the two sets of points are disjoint, then there exists a hyperplane such that all points are correctly classified. However, most practical problems will involve sets of points which cannot be perfectly separated using a hyperplane, so we form an optimization problem whose objective is to minimize some measure of the misclassification. Further, we introduce two parallel-bounding hyperplanes in the middle of which the separating hyperplane lies. We separate the two classes of points by these two hyperplanes, namely

$$\begin{aligned} a^T w - \gamma &= +1 \\ a^T w - \gamma &= -1, \end{aligned} \qquad (2)$$

which bound the classes $\mathcal{A}$ and $\mathcal{B}$, respectively. The capacity of the classifier is the distance between the two bounding hyperplanes given by $\frac{2}{\|w\|}$. Maximizing the capacity has been shown to increase generalization of the classifier to new data points[3].

If the two groups are not linearly separable, we introduce a slack variable $y_i \geq 0$ for each data point. Then, (2) is reformulated as

$$\begin{aligned} a_i^T w - \gamma + y_i &\geq +1, \text{ for points } a_i \text{ in class } \mathcal{A} \\ a_i^T w - \gamma - y_i &\leq -1, \text{ for points } a_i \text{ in class } \mathcal{B}. \end{aligned} \qquad (3)$$

We also want our separating hyperplane to generalize well to additional data points. In order to do this, we need to find the right balance between minimizing the error y of the classifier and maximizing the capacity of the classifier. We see that maximizing this distance is analogous to minimizing the size $l(w)$ of w, the normal of the separating hyperplane. This is performed in the following formulation [20]:

$$\text{(SVM)} \qquad \begin{aligned} &\min_{w,\gamma,y} \; l(w) + \mu e^T y \\ &\text{s.t. } D(Aw - e\gamma) + y \geq e, \quad y \geq 0, \end{aligned}$$

where μ is a weighting parameter and $w \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$, $y \in \mathbb{R}^m$ and $e \in \mathbb{R}^m$ is a vector of ones. When $l(w) = \|w\|_2^2 = w^T w$, (SVM) reduces to a quadratic program. In the case where $l(w) = \|w\|_1 = \sum_{i=1}^n |w_i|$, the above formulation is equivalent to the following linear program:

$$\text{(SVM}_1\text{)} \qquad \begin{aligned} &\min_{w,\gamma,y,t} \; e^T t + \mu e^T y \\ &\text{s.t. } \; D(Aw - e\gamma) + y \geq e, \; y \geq 0, \\ &\qquad t \geq w \geq -t, \end{aligned}$$

where $w, t \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$, $y \in \mathbb{R}^m$. When $l(w) := \frac{\lambda_1}{2}\|w\|_2^2 + \lambda_2\|w\|_1$, the model (SVM) becomes the doubly regularized SVM[21]:

$$\text{(KBP)} \qquad \min_{w,\gamma,y} \frac{\lambda_1}{2}\|w\|_2^2 + \lambda_2\|w\|_1 + \mu e^T y$$
$$\text{s.t.} \quad D(Aw - e\gamma) + y \geq e, \quad y \geq 0.$$

Assume that we now have prior information in the form of a knowledge set, determined by the inequalities $h_i^T z \leq d_i$, $i = 1, 2, \ldots, k$, where h_i's and d_i's are uncertain and they belong to the interval uncertainty set, i.e. $(h_i, d_i) \in [\underline{h}_i, \overline{h}_i] \times [\underline{d}_i, \overline{d}_i]$, $\underline{h}_i, \overline{h}_i \in \mathbb{R}^n$ with $\underline{h}_i < \overline{h}_i$ and $\underline{d}_i, \overline{d}_i \in \mathbb{R}$ with $\underline{d}_i \leq \overline{d}_i$, for $i = 0, 1, \ldots, k$. We further assume that the knowledge set belongs to class $\mathcal{A}$ (Figure 1).

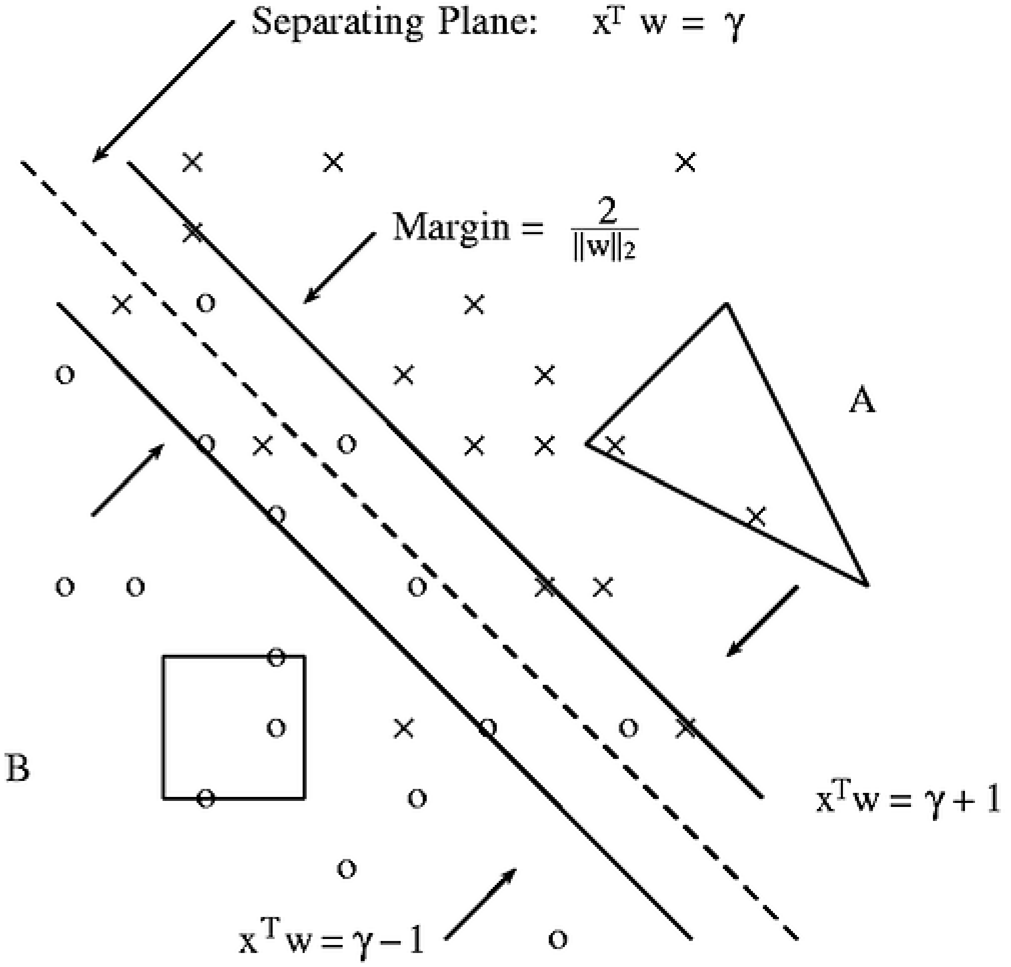

Separating Plane: $x^T w = \gamma$
Margin = $\frac{2}{\|w\|_2}$
A
B
$x^T w = \gamma + 1$
$x^T w = \gamma - 1$

Figure 1. The two bounding planes which define the separating plane with a margin of $\frac{2}{\|w\|_2}$ for groups $\mathcal{A}$ and $\mathcal{B}$. Knowledge sets for classes $\mathcal{A}$ and $\mathcal{B}$ are regions inside the uncertain polyhedral sets

In other words, the uncertain knowledge set $\{z \in \mathbb{R}^n:\ h_i^T z \le d_i,\ i = 1, 2, \dots, k\}$ lies on class $\mathcal{A}$'s side of the bounding hyperplane $w^T z = \gamma + 1$. This is performed in the following knowledge-based SVM model under data uncertainty [6,16]:

$$\begin{aligned}
\text{(KBP)}\qquad & \min_{w,\gamma,y}\ l(w) + \mu e^T y \\
& \text{s.t.}\ \ D(Aw - e\gamma) + y \ge e,\ \ y \ge 0 \\
& \qquad \{z \in \mathbb{R}^n \mid h_i^T z \le d_i,\ i = 1, 2, \dots, k\} \subset \{z \in \mathbb{R}^n:\ \ w^T z \ge \gamma + 1\},
\end{aligned}$$

where $h_i \in \mathbb{R}^n$ and $d_i \in \mathbb{R}^n$ are uncertain for $i = 1, 2, \dots, k$. In particular, when $l(w) := \frac{\lambda_1}{2}\|w\|_2^2 + \lambda_2\|w\|_1$, (KBP) reduces to the doubly regularized knowledge-based SVM under uncertainty:

$$\begin{aligned}
& \min_{w,\gamma,y}\ \frac{\lambda_1}{2}\|w\|_2^2 + \lambda_2\|w\|_1 + \mu e^T y \\
& \text{s.t.}\quad D(Aw - e\gamma) + y \ge e,\quad y \ge 0 \\
& \qquad \{z \in \mathbb{R}^n \mid h_i^T z \le d_i,\ i = 1, 2, \dots, k\} \subset \{z \in \mathbb{R}^n:\ \ w^T z \ge \gamma + 1\},
\end{aligned}$$

where $h_i \in \mathbb{R}^n$ and $d_i \in \mathbb{R}^n$ are uncertain for $i = 1, 2, \dots, k$.

Following robust optimization approach, the robust counterpart 1 of the doubly regularized knowledge-based SVM under uncertainty is a deterministic optimization problem, given by

$$\begin{aligned}
& \min_{w,\gamma,y}\ \frac{\lambda_1}{2}\|w\|_2^2 + \lambda_2\|w\|_1 + \mu e^T y \\
& \text{s.t.}\ \ D(Aw - e\gamma) + y \ge e,\ \ y \ge 0 \\
& \quad \forall (h_i, d_i) \in [\underline{h}_i, \overline{h}_i] \times [\underline{d}_i, \overline{d}_i], \{z \in \mathbb{R}^n \mid h_i^T z \le d_i,\ i = 1, 2, \dots, k\} \subseteq \{z \in \mathbb{R}^n:\ w^T z \ge \gamma + 1\}.
\end{aligned}$$

In the next section, we derive an extension of Farkas' lemma that enables us to convert the above robust counterpart as a convex quadratic program.

Table 1. Performance of the algorithms for three public datasets.

Data set *(m x n)*	Training accuracy	Formulation	Testing accuracy	% No. of selected features
PID (768 x 9)	RK-pq-SVM	0.7802	0.7781	0.4675
	pq-SVM	0.7789	0.7766	0.4750
	L-SVM	0.7711	0.7727	0.4500
WDBC (569 x 30)	RK-pq-SVM	0.9857	0.9830	0.7100
	pq-SVM	0.9825	0.9821	0.7133
	L-SVM	0.9849	0.9821	0.7267
Correlated data (100 x 10)	RK-pq-SVM	0.8500	0.8456	0.4700
	pq-SVM	0.8411	0.8400	0.6600

	L-SVM	0.8367	0.8300	0.4300

3. Robust Farkas' lemma

In this section, we establish an extension of Farkas' lemma 7 to systems involving uncertain linear inequalities with the weighted norm uncertainty. The generalized Farkas' lemma plays a key role in reformulating the doubly regularized knowledge-based SVM with an uncertain knowledge set as a convex quadratic program.

To do this, let us first recall that the usual *p*-norm of $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $p \geq 1$, defined by

$$\|x\|_p = \begin{cases} \left(\sum_{j=1}^{n} |x_j|^p \right)^{\frac{1}{p}}, & \text{if } p \in [1, +\infty), \\ \max_{1 \leq j \leq n} \{|x_j|\}, & \text{if } p = +\infty. \end{cases}$$

The corresponding dual *p*-norm of *x* is given by $\|x\|_p^* := \sup_{\|a\|_p = 1} a^T x = \|x\|_{p^*}$, where p^* satisfies $\frac{1}{p} + \frac{1}{p^*} = 1$. More generally, for a $w = (w_1, \dots, w_n)^T \in \mathbb{R}^n$ with $w_j > 0$, the weighted *p*-norm of *x* is given by

$$\|x\|_{w,p} = \begin{cases} \left(\sum_{j=1}^{n} w_j |x_j|^p \right)^{\frac{1}{p}}, & \text{if } p \in [1, +\infty), \\ \max_{1 \leq j \leq n} \{w_j |x_j|\}, & \text{if } p = +\infty, \end{cases}$$

and the corresponding dual-weighted *p*-norm of *x* is given by $\|x\|_{w,p}^* := \sup_{\|a\|_{w,p} = 1} a^T x = \|Dx\|_{p^*}$, where

$$D = \begin{cases} \operatorname{diag}\left(w_1^{-\frac{1}{p}}, \dots, w_n^{-\frac{1}{p}} \right), & \text{if } p \in (1, +\infty), \\ \operatorname{diag}(w_1^{-1}, \dots, w_n^{-1}), & \text{if } p = \infty. \end{cases}$$

We now present a robust version of the Farkas lemma under the weighted *p*-norm uncertainty.

Theorem 3.1 (Robust Farkas' lemma)

Let $p \in \mathbb{R}$, $p \geq 1$. *Let* $w_i = (w_i^1, \dots, w_i^n)^T \in \mathbb{R}^n$, $i = 1, \dots, k$ *with* $w_i^j > 0$, $c \in \mathbb{R}^n$ *and* $r \in \mathbb{R}$. *Define* $\mathcal{U}_i = \{a_i : \|a_i - \hat{a}_i\|_{w_i,p} \leq \epsilon_i\}$ *and* $\mathcal{V}_i = \{\beta_i : |\beta_i - \hat{\beta}_i| \leq \delta_i\}$, *where* $\hat{a}_i \in \mathbb{R}^n$, $\hat{\beta}_i \in \mathbb{R}$, $\varepsilon_i, \delta_i \in \mathbb{R}_+$ *for* $i = 0, 1, \dots, k$.*Suppose that one of the following two conditions holds*:

1. $p = 1$ *or* $p = +\infty$

2. $p \in (1, +\infty)$ *and there exists* $x_0 \in \mathbb{R}^n$ *such that* $a_i^T x_0 < \beta_i$, *for all* $(a_i, \beta_i) \in \mathcal{U}_i \times \mathcal{V}_i$.

Then, the following statements are equivalent:

i. $\forall (a_i, \beta_i) \in \mathcal{U}_i \times \mathcal{V}_i,\ a_i^T x \leq \beta_i,\ i = 1, \ldots, k \Rightarrow c^T x \geq r$.

ii. $(\exists \lambda \in \mathbb{R}_+^k, u_i \in \mathbb{R}^n\ with\ \|u_i\|_p \leq 1)$
$(c + \sum_{i=1}^k \lambda_i \hat{a}_i) + \sum_{i=1}^k \lambda_i \epsilon_i D_i u_i = 0\ and\ r + \sum_{i=1}^k \lambda_i (\hat{\beta}_i - \delta_i) \leq 0$.

Proof [(i) ⇒ (ii)]

Define $g_i(x) = \sup_{(a_i, \beta_i) \in \mathcal{U}_i \times \mathcal{V}_i} \{a_i^T x - \beta_i\}$, $i = 1, \ldots, k$. Then,

$$\begin{aligned} g_i(x) &= \sup\{a_i^T x - \beta_i : a_i = \hat{a}_i + u_i, \|u_i\|_{w_i,p} \leq \epsilon_i,\ \hat{\beta}_i - \delta_i \leq \beta_i \leq \hat{\beta}_i - \delta_i\} \\ &= \hat{a}_i^T x + \epsilon_i \|x\|_{w_i,p}^* - (\hat{\beta}_i - \delta_i) \\ &= \hat{a}_i^T x + \epsilon_i \|D_i x\|_{p^*} - (\hat{\beta}_i - \delta_i). \end{aligned}$$

Then, (i) can be equivalently rewritten as

$$g_i(x) \leq 0, \quad i = 1, \ldots, k \Rightarrow c^T x - r \geq 0.$$

If the assumption (1) holds, then each g_i is a polyhedral function. On the other hand, if the assumption (2) holds, then the Slater condition (i.e. $\{x : g_i(x) < 0, i = 1, \ldots, k\} \neq \emptyset$) is verified. In both cases, the classical Farkas' lemma shows us that there exists a $\lambda_i \geq 0$ such that

$$(c^T x - r) + \sum_{i=1}^k \lambda_i g_i(x) \geq 0, \quad \forall x \in \mathbb{R}^n.$$

This implies that

$$\left(c + \sum_{i=1}^k \lambda_i \hat{a}_i\right)^T x + \sum_{i=1}^k \lambda_i \epsilon_i \|D_i x\|_{p^*} \geq 0, \quad \forall x \in \mathbb{R}^n \text{ and } -r - \sum_{i=1}^k \lambda_i (\hat{\beta}_i - \delta_i) \geq 0.$$

The first condition is equivalent to the inclusion

$$0 \in \left(c + \sum_{i=1}^{k} \lambda_i \hat{a}_i\right) + \sum_{i=1}^{k} \lambda_i \epsilon_i D_i \partial\,(\|\cdot\|_{p^*})(0),$$

where ∂ is the standard convex subdifferential [18]. This means that (due to the fact that $\partial\,(\|\cdot\|_{p^*})(0) = \{x : \|x\|_p \leq 1\}$) there exists a $u_i \in \mathbb{R}^n$ with $\|u_i\|_p \leq 1$ such that

$$\left(c + \sum_{i=1}^{k} \lambda_i \hat{a}_i\right) + \sum_{i=1}^{k} \lambda_i \epsilon_i D_i u_i = 0.$$

Thus, statement (ii) holds.

[(ii) ⇒ (i)] Take $x \in \mathbb{R}^n$ such that $a_i^T x \leq \beta_i,\ i = 1, \ldots, k$, $\forall (a_i, \beta_i) \in \mathcal{U}_i \times \mathcal{V}_i$. Then,

$$\hat{a}_i^T x + \epsilon_i \|D_i x\|_{p^*} - (\hat{\beta}_i - \delta_i) = g_i(x) \leq 0, \quad i = 1, \ldots, k.$$

This together with statement (ii) implies that

$$\begin{aligned} c^T x - r &\geq c^T x - r + \sum_{i=1}^{k} \lambda_i \big(\hat{a}_i^T x + \epsilon_i \|D_i x\|_{p^*} - (\hat{\beta}_i - \delta_i)\big) \\ &= \left(c + \sum_{i=1}^{k} \lambda_i \hat{a}_i\right)^T x + \left(\sum_{i=1}^{k} \lambda_i \epsilon_i \|D_i x\|_{p^*}\right) - \left(r + \sum_{i=1}^{k} \lambda_i (\hat{\beta}_i - \delta_i)\right) \\ &= -\sum_{i=1}^{k} \lambda_i \epsilon_i D_i u_i^T x + \left(\sum_{i=1}^{k} \lambda_i \epsilon_i \|D_i x\|_{p^*}\right) - \left(r + \sum_{i=1}^{k} \lambda_i (\hat{\beta}_i - \delta_i)\right) \\ &= \sum_{i=1}^{k} \lambda_i \epsilon_i (\|D_i x\|_{p^*} - u_i^T D_i x) - \left(r + \sum_{i=1}^{k} \lambda_i (\hat{\beta}_i - \delta_i)\right) \geq 0. \end{aligned}$$

Thus, statement (i) holds.▪

Remark 3.1

It should be noted that, in the special case of $p = +\infty$, $\varepsilon_i = 0$ and $\delta_i = 0$, $i = 1, \ldots, k$, Theorem 3.1 reduces to the celebrated Farkas lemma (cf. 7). Various extensions of the Farkas lemma can be found in [8,10].

We now examine the robust Farkas lemma of Theorem 3.1 in the case of interval uncertainty. This case will enable us to reformulate the knowledge-based SVM with uncertain knowledge sets as a standard quadratic program in the next section. Moreover, the interval uncertainty is the simplest and most commonly used uncertainty in robust optimization [1].

We see that the interval uncertainty case can be obtained as a special case of the weighted ∞-norm uncertainty. To see this, consider $\underline{h} = (\underline{h}^1, \ldots, \underline{h}^n)^T \in \mathbb{R}^n$, $\overline{h} = (\overline{h}^1, \ldots, \overline{h}^n)^T \in \mathbb{R}^n$, $\overline{h} = (\overline{h}^1, \ldots, \overline{h}^n)^T \in \mathbb{R}^n$ with $\underline{h} < \overline{h}$. Let $\epsilon = \max_{1 \leq j \leq n}\{\overline{h}^j - \underline{h}^j\} > 0$. Let $w = (w^1, \ldots, w^n)$ with $w^j = \frac{2\epsilon}{\overline{h}^j - \underline{h}^j} > 0$, $\hat{h} = \frac{\underline{h} + \overline{h}}{2}$. Then, we obtain

$$[\underline{h}, \overline{h}] = \left\{(h^1, \ldots, h^n) : \max_{1 \leq j \leq n} w^j \, |h^j - \hat{h}^j| \leq \epsilon\} = \{h : \|h - \hat{h}\|_{w,p} \leq \epsilon\right\}.$$

Proposition 3.1

Let $\underline{h}_i, \overline{h}_i \in \mathbb{R}^n$ *with* $\underline{h}_i < \overline{h}_i$ *and* $\underline{d}_i, \overline{d}_i \in \mathbb{R}$ *with* $\underline{d}_i \leq \overline{d}_i$, *for i* = 0, 1, … , *k. Then, the following statements are equivalent*:

i. $h_i^T x \leq d_i, \ \forall (h_i, d_i) \in [\underline{h}_i, \overline{h}_i] \times [\underline{d}_i, \overline{d}_i], \ i = 1, \ldots, k \Rightarrow w^T x \geq r.$

ii. $(\exists \lambda \in \mathbb{R}^k_+) \ w + \sum_{i=1}^k \lambda_i \overline{h}_i \geq 0, \ w + \sum_{i=1}^k \lambda_i \underline{h}_i \leq 0 \, and \, r + \sum_{i=1}^k \lambda_i \underline{d}_i \leq 0.$

Proof

Let $p = +\infty$, $\epsilon_i = \max_{1 \leq j \leq n}\{\overline{h}_i^j - \underline{h}_i^j\} > 0$. Let $w_i = (w_i^1, \ldots, w_i^n)$ with $w_i^j = \frac{2\epsilon_i}{\overline{h}_i^j - \underline{h}_i^j} > 0$, $\hat{h}_i = \frac{\underline{h}_i + \overline{h}_i}{2}$. Then, we have $[\underline{h}, \overline{h}] = \{h : \|h - \hat{h}\|_{w_i,p} \leq \epsilon_i\}$. So, Theorem 3.1 shows us that there exist a $\lambda \in \mathbb{R}^k_+, \ u_i \in \mathbb{R}^n \text{ with } ||u_i||_\infty \leq 1$such that $w + \sum_{i=1}^k \lambda_i \hat{h}_i + \sum_{i=1}^k \lambda_i \epsilon_i D_i u_i = 0$ and $r + \sum_{i=1}^k \lambda_i \underline{d}_i \leq 0$, where $D_i = \mathrm{diag}((w_i^1)^{-1}, \ldots, (w_i^n)^{-1})$. Note that

$$\epsilon_i D_i = \mathrm{diag}\left(\frac{\overline{h}_i^1 - \underline{h}_i^1}{2}, \ldots, \frac{\overline{h}_i^n - \underline{h}_i^n}{2}\right).$$

So, the first condition can be equivalently rewritten as

$$\|u_i\|_\infty \leq 1 \text{ and } w + \sum_{i=1}^k \lambda_i \left(\frac{\underline{h}_i + \overline{h}_i}{2} + \mathrm{diag}\left(\frac{\overline{h}_i^1 - \underline{h}_i^1}{2}, \ldots, \frac{\overline{h}_i^n - \underline{h}_i^n}{2}\right) u_i\right) = 0.$$

This is, in turn, equivalent to

$$w + \sum_{i=1}^k \lambda_i \overline{h}_i \geq 0 \quad \text{and} \quad w + \sum_{i=1}^k \lambda_i \underline{h}_i \leq 0.$$

▪

4. Knowledge-based SVMs under uncertainty

In this section, we use Proposition 3.1 to derive an equivalent quadratic program for the uncertain (KBP) with the interval knowledge data uncertainty, extending the recent doubly regularized SVM model [4,5].

Let $\{z \in \mathbb{R}^n | h_i^T z \leq d_i, i = 1, 2 \ldots, k\}$ be our uncertain knowledge set for points in class $\mathcal{A}$. We would like the robust counterpart of the knowledge set to be in the region $w^T z \geq \gamma + 1$. Thus, our robust counterpart of the set containment constraint in (KBP) is

$$\forall (h_i, d_i) \in [\underline{h}_i, \overline{h}_i] \times [\underline{d}_i, \overline{d}_i],\ \{x : h_i^T z \leq d_i, i = 1, \ldots, k\} \subseteq \{w^T z \geq \gamma + 1\}. \qquad (4)$$

Now Proposition 3.1 shows that

$$\forall (h_i, d_i) \in [\underline{h}_i, \overline{h}_i] \times [\underline{d}_i, \overline{d}_i], \{z : h_i^T z \leq d_i, i = 1, \ldots, k\} \subseteq \{w^T z \geq \gamma + 1\}$$

$$\Leftrightarrow \exists u_i \geq 0 \ \text{ s.t. } \begin{cases} w + \sum_{i=1}^{k} u_i \underline{h}_i \leq 0, \\ w + \sum_{i=1}^{k} u_i \overline{h}_i \geq 0, \\ \gamma + 1 + \sum_{i=1}^{k} u_i \underline{d}_i \leq 0. \end{cases} \qquad (5)$$

Incorporating these constraints into the doubly regularized SVM formulation, we obtain the following robust knowledge-based doubly regularized SVM problem:

$$\min_{(w,\xi,\gamma,u_i) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^k} \quad \frac{\lambda_1}{2} \|w\|_2^2 + \lambda_2 \|w\|_1 + e_m^T \xi$$

$$\text{s.t.} \quad \begin{cases} D(Aw - \gamma e_m) + \xi \geq e_m, \\ -w - \sum_{i=1}^{k} u_i \underline{h}_i \geq 0, \\ w + \sum_{i=1}^{k} u_i \overline{h}_i \geq 0, \\ -(\gamma + 1) - \sum_{i=1}^{k} u_i \underline{d}_i \geq 0, \\ \xi \geq 0, \ u_i \geq 0, \quad i = 1, \ldots, k. \end{cases}$$

Adding slack variables $\zeta_1, \zeta_2 \in \mathbb{R}_+^n$ and $\beta \geq 0$ and minimizing their L_1 norm, the model becomes

$$
(P_0) \quad \min_{(w,\xi,\gamma,u_i)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\times\mathbb{R}} \quad \frac{\lambda_1}{2}\|w\|_2^2+\lambda_2\|w\|_1+e_m^T\xi+e_n^T(\zeta_1+\zeta_2)+\beta
$$

$$
\text{s.t.} \quad \begin{cases} D(Aw-\gamma e_m)+\xi\geq e_m, \\ -w-\sum_{i=1}^{k}u_i\underline{h}_i+\zeta_1\geq 0, \\ w+\sum_{i=1}^{k}u_i\overline{h}_i+\zeta_2\geq 0, \\ -(\gamma+1)-\sum_{i=1}^{k}u_i\underline{d}_i+\beta\geq 0, \\ \xi\geq 0,\zeta_1,\zeta_2\geq 0,\beta\geq 0, \\ u_i\geq 0, \qquad i=1,\ldots,k. \end{cases}
$$

We see that in the absence of knowledge sets in (P_0), where $w=0$ and $\gamma=-1$, our robust knowledge-based doubly-regularized SVM problem reduces to

$$
\min_{(w,\xi,\gamma,u_i)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\times\mathbb{R}} \quad \frac{\lambda_1}{2}\|w\|_2^2+\lambda_2\|w\|_1+e_m^T\xi+e_n^T(\zeta_1+\zeta_2)+\beta
$$

$$
\text{s.t.} \quad \begin{cases} D(Aw-\gamma e_m)+\xi\geq e_m, \\ -\sum_{i=1}^{k}u_i\underline{h}_i+\zeta_1\geq 0, \\ \sum_{i=1}^{k}u_i\overline{h}_i+\zeta_2\geq 0, \\ -\sum_{i=1}^{k}u_i\underline{d}_i+\beta\geq 0, \\ \xi\geq 0,\zeta_1,\zeta_2\geq 0,\beta\geq 0, \\ u_i\geq 0, \quad i=1,\ldots,k, \end{cases}
$$

which is equivalent to the following doubly regularized SVM problem proposed in 5:

$$
\min_{(w,\xi,\gamma)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}} \quad \frac{\lambda_1}{2}\|w\|_2^2+\lambda_2\|w\|_1+e_m^T\xi
$$

$$
\text{s.t.} \quad \begin{cases} D(Aw-\gamma e_m)+\xi\geq e_m. \\ \xi\geq 0 \end{cases}
$$

Therefore, the model problem (P_0) is an extension of the SVM model considered in 5 by incorporating uncertain knowledge sets.

Let $w=p-q$, where $p=(p_1,\ldots,p_n)$ and $q=(q_1,\ldots,q_n)$, be defined by

$$p_i = (w_i)_+ = \begin{cases} 0, & w_i \le 0, \\ w_i, & w_i > 0 \end{cases} \quad \text{and} \quad q = (w_i)_- = \begin{cases} 0, & w_i \ge 0, \\ -w_i, & w_i < 0. \end{cases}$$

Then,

$$\|w\|_2^2 = \|p - q\|_2^2 = \|p\|_2^2 + \|q\|_2^2 \quad \text{and} \quad \|w\|_1 = e_n^T(p+q).$$

So, the robust knowledge-based doubly regularized SVM problem (P_0) can be rewritten as

$$\min_{(p,q,\xi,\gamma,u_i) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}} \quad \frac{\lambda_1}{2}(\|p\|_2^2 + \|q\|_2^2) + \lambda_2 e_n^T(p+q) + e_m^T\xi + e_n^T(\zeta_1 + \zeta_2) + \beta$$

$$\text{s.t.} \quad \begin{cases} D\big(A(p-q) - \gamma e_m\big) + \xi \ge e_m, \\ -p + q - \sum_{i=1}^{k} u_i \underline{h}_i + \zeta_1 \ge 0, \\ p - q + \sum_{i=1}^{k} u_i \overline{h}_i + \zeta_2 \ge 0, \\ -(\gamma+1) - \sum_{i=1}^{k} u_i \underline{d}_i + \beta \ge 0, \\ p, q \ge 0, \ \xi \ge 0, \zeta_1, \zeta_2 \ge 0, \beta \ge 0, \\ u_i \ge 0, \quad i = 1, \ldots, k. \end{cases}$$

We now further simplify this model to a form of quadratic program in the matrix form. To do this, denote

$$H_1 = (\underline{h}_1, \ldots, \underline{h}_k) \in \mathbb{R}^{n\times k}, \quad H_2 = (\overline{h}_1, \ldots, \overline{h}_k) \in \mathbb{R}^{n\times k},$$

$$d = (\underline{d}_1, \ldots, \underline{d}_k)^T \in \mathbb{R}^k, \quad u = (u_1, \ldots, u_k)^T \in \mathbb{R}^k,$$

$$y = \begin{pmatrix} p \\ q \\ u \end{pmatrix} \in \mathbb{R}^{2n+k}, \quad b = \lambda_2 \begin{pmatrix} e_n \\ e_n \\ 0 \end{pmatrix} \in \mathbb{R}^{2n+k},$$

$$v = \begin{pmatrix} \xi \\ \zeta_1 \\ \zeta_2 \\ \beta \end{pmatrix} \in \mathbb{R}^{m+2n+1}, \quad \hat{e} = \begin{pmatrix} e_m \\ 0 \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{m+2n+1}, \qquad (6)$$

$$C = \lambda_1 \begin{pmatrix} I_{n\times n} & 0_{n\times n} & 0_{n\times k} \\ 0_{n\times n} & I_{n\times n} & 0_{n\times k} \\ 0_{k\times n} & 0_{k\times n} & 0_{k\times k} \end{pmatrix} \in \mathbb{R}^{(2n+k)\times(2n+k)}, \tag{7}$$

$$\hat{D} = \begin{pmatrix} D & 0_{m\times(2n+1)} \\ 0_{(2n+1)\times m} & I_{(2n+1)\times(2n+1)} \end{pmatrix} \in \mathbb{R}^{(m+2n+1)\times(m+2n+1)} \tag{8}$$

and

$$\hat{A} = \begin{pmatrix} A & -A & 0_{m\times k} \\ -I_{n\times n} & I_{n\times n} & -H_1 \\ I_{n\times n} & -I_{n\times n} & H_2 \\ 0 & 0 & -d^T \end{pmatrix} \in \mathbb{R}^{(m+2n+1)\times(2n+k)}. \tag{9}$$

Then, we can write the robust knowledge-based regularized SVM problem (P_0) into the following matrix form:

$$\begin{aligned} \min_{(y,v,\gamma)\in\mathbb{R}^{2n+k}\times\mathbb{R}^{m+2n+1}\times\mathbb{R}} \quad & \frac{1}{2}y^TCy + b^Ty + e_{m+2n+1}^T v \\ \text{s.t.} \quad & \hat{D}(\hat{A}y - \gamma\hat{e}) + v \geq \hat{e} \\ & y \geq 0, v \geq 0. \end{aligned}$$

Using a similar idea to that of Mangasarian for the LSVM 16, we replace b^Ty by y^Ty and $e_{m+2n+1}^T v$ by $\frac{1}{2}v^Tv$. This allows us to remove the non-negative constraints $y \geq 0$ and $v \geq 0$. Moreover, we also append an additional $\frac{\gamma^2}{2}$ to the objective function as done in 17. This in effect maximizes the margin between the parallel separating planes. These modifications give rise to the following optimization problem:

$$\begin{aligned} (P) \quad \min_{(y,v,\gamma)\in\mathbb{R}^{2n+k}\times\mathbb{R}^{m+2n+1}\times\mathbb{R}} \quad & \frac{1}{2}y^T(C+\mu I)y + \frac{1}{2}\gamma^2 + \frac{1}{2}\|v\|_2^2 \\ \text{s.t.} \quad & \hat{D}(\hat{A}y - \gamma\hat{e}) + v \geq \hat{e}, \end{aligned}$$

where $\mu \in \mathbb{R}$ is an additional tuning parameter.

5. Duality and algorithm

Jump to section

In this section, we present an algorithm for finding a solution of (P) by solving its Lagrangian dual problem. We also provide a proof for the convergence of the algorithm.

To formulate its Lagrangian dual, we define the Lagrangian as follows:

$$L(y,v,\gamma,z)=\frac{1}{2}y^T(C+\mu I)y+\frac{1}{2}\gamma^2+\frac{1}{2}\|v\|_2^2-z^T(\hat{D}\hat{A}y-\gamma\hat{D}\hat{e}+v-\hat{e}).$$

Then, the Lagrangian dual problem becomes

$$\begin{aligned}\max_{y,v,\gamma\in\mathbb{R},z\geq 0}\quad & L(y,\gamma,z)\\ \text{s.t.}\quad & \frac{\partial L}{\partial y}(y,v,\gamma,z)=0,\ \frac{\partial L}{\partial v}(y,v,\gamma,z)=0,\ \frac{\partial L}{\partial \gamma}(y,v,\gamma,z)=0.\end{aligned}$$

This can be expressed as

$$\begin{aligned}\max_{y\in\mathbb{R}^{n+2k},\,v\in\mathbb{R}^{m+2n+1},\,\gamma\in\mathbb{R},\,z\geq 0}\quad & \frac{1}{2}y^T(C+\mu I)y+\frac{1}{2}\gamma^2+\frac{1}{2}\|v\|_2^2-z^T(\hat{D}\hat{A}y-\gamma\hat{D}\hat{e}+v-\hat{e})\\ \text{s.t.}\quad & (C+\mu I)y-(\hat{D}\hat{A})^Tz=0,\ v-z=0,\ \gamma+(\hat{D}\hat{e})^Tz=0.\end{aligned}$$

Solving the constraints gives us that

$$y=(C+\mu I)^{-1}(\hat{D}\hat{A})^Tz,\ v=z\quad\text{and}\quad\gamma=-(\hat{D}\hat{e})^Tz.$$

Substituting these two relations into the Lagrangian dual, we get

$$\begin{aligned}\max_{z\in\mathbb{R}^{m+2n+1}}\quad & -\frac{1}{2}z^T\left(\hat{D}\hat{A}(C+\mu I)^{-1}(\hat{D}\hat{A})^T\right)z-\frac{1}{2}\left((\hat{D}\hat{e})^Tz\right)^2-\frac{1}{2}\|z\|^2+\hat{e}^Tz\\ \text{s.t.}\quad & z\geq 0.\end{aligned}$$

Note that $((\hat{D}\hat{e})^Tz)^2=(z^T(\hat{D}\hat{e}))(\hat{D}\hat{e})^Tz)=z^T(\hat{D}\hat{e}(\hat{D}\hat{e})^T)z$, and so, the dual problem can be rewritten as

$$\begin{aligned}\max_{z\in\mathbb{R}^{m+2n+1}}\quad & -\frac{1}{2}z^T\left(I+\hat{D}\hat{A}(C+\mu I)^{-1}(\hat{D}\hat{A})^T+\hat{D}\hat{e}(\hat{D}\hat{e})^T\right)z+\hat{e}^Tz\\ \text{s.t.}\quad & z\geq 0.\end{aligned}$$

Letting $Q:=I+\hat{D}\hat{A}(C+\mu I)^{-1}(\hat{D}\hat{A})^T+\hat{D}\hat{e}(\hat{D}\hat{e})^T$, we see that Q is positive definite as for each $x\in\mathbb{R}^{m+2n+1}$,

$$x^TQx=x^T\left(I+\hat{D}\hat{A}(C+\mu I)^{-1}(\hat{D}\hat{A})^T+\hat{D}\hat{e}(\hat{D}\hat{e})^T\right)x\geq\|x\|^2.$$

This shows that the dual is equivalent to the following strictly concave quadratic maximization problem with non-negativity constraints:

$$(D) \ \max_{z \geq 0} -\frac{1}{2} z^T Q z + \hat{e}^T z.$$

The following theorem presents the duality relationship between (P) and (D).

Theorem 5.1

Let $(y, v, \gamma) \in \mathbb{R}^{2n+k} \times \mathbb{R}^{m+2n+1} \times \mathbb{R}$ *and* $z \in \mathbb{R}^{m+2n+1}$*. Then z is a solution of* (D) *if and only if* $(C + \mu I)^{-1}(\hat{D}\hat{A})^T z, z, -(\hat{D}\hat{e})^T z)$ *is a solution of* (P)*. Moreover, we have*

$$\min(P) = \max(D).$$

Proof

Clearly,

$$\{(y, \gamma, v) : \hat{D}(\hat{A}y - \gamma\hat{e}) + v > \hat{e}\} \neq \emptyset.$$

So, it follows from the Lagrangian duality theorem that

$$\min(P) = \max(D) = \max_{y,\gamma \in \mathbb{R}, z \geq 0} \left\{ L(y, v, \gamma, z) : \frac{\partial L}{\partial y}(y, v, \gamma, z) = 0, \right.$$
$$\left. \frac{\partial L}{\partial v}(y, v, \gamma, z) = 0, \frac{\partial L}{\partial \gamma}(y, v, \gamma, z) = 0 \right\}.$$

Note that $\frac{\partial L}{\partial y}(y, v, \gamma, z) = (C + \mu I)\, y - (\hat{D}\hat{A})^T z$, $\frac{\partial L}{\partial v}(y, v, \gamma, z) = v - z$ and $\frac{\partial L}{\partial \gamma}(y, v, \gamma, z) = \gamma + (\hat{D}\hat{e})^T z.$So, the conclusion follows.▪

Now, to solve the dual problem, let us look at its optimality condition, which is a simple nonlinear complementary problem $0 \leq z \perp Qz - \hat{e} \geq 0$. By using the following elementary equivalence

$$0 \leq a \perp b \geq 0 \Leftrightarrow b = (b - \alpha a)_+, \quad \alpha > 0,$$

the optimality condition reduces to

$$Qz - \hat{e} = \big((Qz - \hat{e}) - \alpha z\big)_+.$$

This leads us to the following simple iterative fixed point algorithm:

$$z^{i+1} = Q^{-1}(\hat{e} + \big((Qz^i - e) - \alpha z\big)_+),$$

where α is a real number satisfying $0 < \alpha < 2$.

To summarize, we formulate the pseudo-algorithm (Algorithm 1) as follows.

Algorithm 1

Tuning procedure

Construct a grid with each grid point corresponding to the pair

$$(\mu_i, (\lambda_1)_j) = (2^i, 2^j), \ \ (i, j) \in \{-5, -4, \ldots, 10\}.$$

Select a tuning parameter and tuning set.

Inner problem

For the selected tuning parameter μ, λ_1, determine the matrix

$$Q = I + \hat{D}\hat{A}(C + \mu I)^{-1}(\hat{D}\hat{A})^T + \hat{D}\hat{e}(\hat{D}\hat{e})^T,$$

where C is defined as in (4.7), $\hat{D}$ is defined as in (4.8), $\hat{A}$ is defined as in (4.9) and $\hat{e}$ is defined as in (4.6). Solve the inner problem with the following steps.

```
Step 1   let $z^0 = Q^{-1}\hat{e}$, set it = 0 and $i = 0$

Step 2   $z_{old} = z^0 + \hat{e}$

Step 3   While it < maxiter and $\|z_{old} - z^i\| > tol$
    $z_{old} = z^i$
    $z^{i+1} = Q^{-1}(\hat{e} + ((Qz^i - \hat{e};) - \alpha z^i))_+)$
    it = it + 1 and $i = i + 1$;
end

Step 4   Calculate $y = (C + \mu I)^{-1}(\hat{D}\hat{A})^T z$ and $\gamma = -(\hat{D}\hat{e})^T z$. Identify $p, q$ by $y = \begin{pmatrix} p \\ q \\ u \end{pmatrix}$.
Output $\gamma$ and$w = p - q$ and record the test accuracy, CPU time and the useful features by
removing all features corresponding to weights satisfying $|w_i|/\|w\|_\infty < 0.25$.
```

Updating the tuning parameter

Using the 10-fold cross-validation and update the tuning parameter.

Output

Determine the optimal tuning parameters by choosing the highest average testing accuracy. Then output the corresponding testing accuracy, training accuracy, average CPU time and average selected features.

Now, we present the convergence of our algorithm.

Theorem 5.2

Let $0 < \alpha < 2$ *and let* μ, λ_1 *be arbitrary tuning parameters. Let* $\{z^i\}_{i=0}^{\infty}$ *be a sequence generated by the inner problem. Then* z^i*converges to a unique solution z of the dual problem* (D).

Proof

As D is a strictly concave maximization problem, it has a unique solution provided the solution set is non-empty. Now, denote the unique solution by z. Let $a = Qz$ and $a^i = Qz^i$ for each $i = 0, 1, \ldots$. To show $z^i \rightarrow z$, we only need to show $a^i \rightarrow a$ as Q is positive definite (and so, invertible). By the optimality condition, we see that

$$Qz - \hat{e} = \left((Qz - \hat{e}) - \alpha z\right)_+,$$

and hence

$$a = \hat{e} + \left(a - \hat{e} - \alpha Q^{-1} a\right)_+.$$

On the other hand, by our algorithm, $z^{i+1} = Q^{-1}(\hat{e} + ((Qz^i - \hat{e}) - \alpha z^i)_+)$.
So, $a^{i+1} = \hat{e} + (u^i - \hat{e} - \alpha Q^{-1} a^i)_+$,

$$\|a^{i+1} - a\| = \|(\left(a^i - \hat{e} - \alpha Q^{-1} a^i\right)_+ - \left((a - \hat{e}) - \alpha Q^{-1} a\right)_+\|.$$

Now, using the projection theorem, which states that the distance between any two points is not less than the distance between their projections on any convex set (here is the nonnegative orthant), the above relation gives us

$$\|a^{i+1} - a\| \leq \|(I - \alpha Q^{-1})(a^i - a)\| \leq \|I - \alpha Q^{-1}\| \, \|a^i - a\|.$$

To finish the proof, it suffices to show that $\|I - \alpha Q^{-1}\| < 1$. To see this, note that $0 < \alpha < 2$ and $x^T Qx \geq \|x\|^2$ for each x. So, for each z, we have $\|z\| \, \|Q^{-1} z\| \geq z^T Q^{-1} z \geq \|Q^{-1} z\|^2$. Thus, $\|Q^{-1}\| \leq 1$. So, whenever $\alpha \in (0, 2)$, $\|I - \alpha Q^{-1}\| < 1$. Hence, the conclusion follows.▪

6. Computational experiments

In this section, we provide details on the computer implementation of our proposed algorithm.

Datasets

To conduct the analysis, three publicly available datasets were utilized. These datasets are accessible via the Wisconsin machine learning website: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

For the reader's interest, a short summary of each dataset is included below.

- Wisconsin Breast Cancer dataset The Wisconsin Breast Cancer dataset (WDBC) consists of 30 real-valued features, constructed from 10 characteristics within the lump of 569 women with suspected breast cancer.

- Pima Indians dataset The Pima Indians dataset (PID) consists of 768 observations, each with eight features describing attributes such as blood pressure and body mass index of both healthy patients and those displaying signs of diabetes among the Pima Indian population.

- Correlated data This dataset was constructed using MATLAB, and consists of 10 features; the first five of which are highly correlated. These correlated features are referred to as signal variables, whereas the remaining features are regarded as noisy, irrelevant variables. The class +1 follows a normal distribution with mean $\mu_+ = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)^T$ and with a covariance matrix

  $$\Sigma = \begin{pmatrix} \Sigma^* & 0_{5\times 5} \\ 0_{5\times 5} & I_{5\times 5} \end{pmatrix},$$

  where Σ^* is a (5×5) matrix such that each diagonal element is 1 and each off-diagonal element is 0.8. The class −1 is also normally distributed with the same covariance matrix but with mean $\mu_- = (-1, -1, -1, -1, -1, 0, 0, 0, 0, 0)^T$.

Methods

- Robust knowledge-based pq-SVM (RK-pq-SVM) method (Algorithm 1): In particular, the uncertain knowledge set is generated by the following procedures: we first took a small part of the data in the given dataset to form a preliminary knowledge set. Then, we randomly generated 100 independent copies from this preliminary knowledge set by adding Gaussian noises. Then, the lower-bound ($\underline{h}_i$ and $\underline{d}_i$) and upper-bound ($\overline{h}_i$ and $\overline{d}_i$) of the uncertain knowledge set were determined as the smallest lower-bound and biggest upper-bound of these 100 copies.

- The pq-SVM method (Code was based on 5).

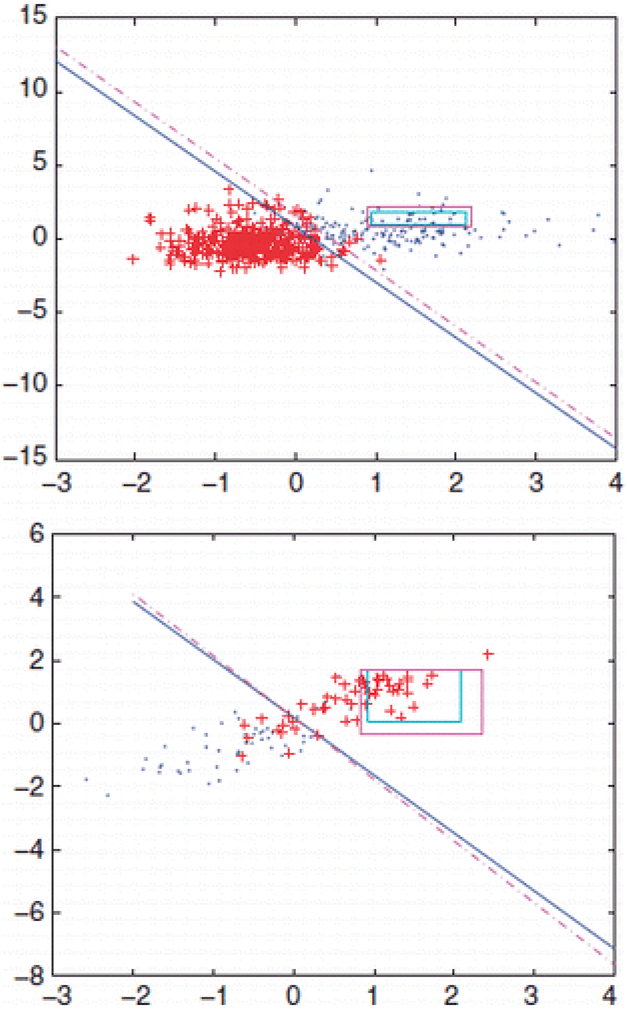
- The Lagrangian-SVM (L-SVM) method (Code was based on Mangasarian 17).

Comparison

From Table 1, we can see that the RK-pq-SVM slightly outperforms pq-SVM and Lagrangian-SVM in terms of both training accuracy and testing. In terms of selecting the fewer features, the L-SVM is comparable with the RK-pq-SVM and pq-SVM, and RK-pq-SVM slightly outperforms pq-SVM.

6.1. Visualization of the results

In this subsection, we present graphs to visualize how incorporating robust knowledge sets affects the resulting separation planes. To see the results, we pick the first two features in each of the two datasets and plot the two classes of data sets, separation hyperplanes and the uncertain polyhedral knowledge sets. The results for the datasets MDBC and Correlated_Data are given below.



7. Conclusion and future research

In this article, we have shown how data uncertainty in knowledge sets can be treated in SVM classification by employing robust optimization. We examined knowledge-based SVMs within the framework of robust optimization that incorporates prior knowledge in the form of uncertain linear constraints. By using a new robust version of Farkas' lemma under uncertainty, we reformulated the knowledge-based SVM problem as a standard quadratic optimization problem. A solution of the reformulated problem was then obtained using the Lagrangian duality scheme and a fixed point iterative algorithm. We also proved the convergence of the algorithm. We finally provided some preliminary results on the implementation of our numerical scheme. Our approach raises some interesting questions for further research.

For instance, it is known that the use of nonlinear kernels in SVM formulations is generally effective in knowledge-based classification. It would be of interest to extend our approach to solve classification problems with positive semidefinite nonlinear kernels. On the other hand, an efficient construction of uncertainty sets is a key modelling issue in the area of robust optimization. Consequently, it would be beneficial from the point of view of practical applications to study robust optimization models with other broad classes of uncertainty sets, such as the ellipsoidal uncertainty, and to examine efficient ways of constructing these uncertainty sets for SVM classification. These issues will be investigated in a forthcoming study.

Acknowledgements

This research was partially supported by a grant from the Australian Research Council.

References

1. Ben-Tal, A, Ghaoui, LE and Nemirovski, A. 2009. *Robust Optimization, Princeton Series in Applied Mathematics*, Princeton, NJ: Princeton University Press.

2. Ben-Tal, A and Nemirovski, A. 2002. Robust optimization – Methodology and applications. *Math. Program Ser. B*, 92: 453–480.

3. Burgess, CJC. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2: 121–167.

4. Cysiqued, LA, Murray, JM, Dunbar, M, Jeyakumar, V and Brew, BJ. 2010. A screening algorithm for HIV-associated neurocognitive disorders. *HIV Med.*, 11: 642–649.

5. Dunbar, M, Murray, JM, Cysiqued, LA, Brew, BJ and Jeyakumar, V. 2010. Simultaneous classification and feature selection via convex quadratic programming with application to HIV-associated neurocognitive disorder assessment. *Eur. J. Oper. Res.*, 206: 470–478.

6. Fung, G, Mangasarian, OL and Shavlik, J. 2003. “Knowledge-based support vector machine classifiers”. In *Neural Information Processing Systems. Vol. 15*, Edited by: Thurn, S, Becker, S and Obermayer, K. 521–528. Cambridge, MA: MIT Press.

7. Jeyakumar, V. 2001. “Farkas lemma: Generalizations”. In *Encyclopedia of Optimization*, Edited by: Floudas, CA and Pardalos, PM. 87–91. Boston, , USA: Vol. 2, Kluwer Academic Publishers.

8. Jeyakumar, V. 2003. Characterizing set containments involving infinite convex constraints and reverse-convex constraints.*SIAM J. Optim.*, 13: 947–959.

9. Jeyakumar, V and Li, G. 2010. Strong duality in robust convex programming: Complete characterizations. *SIAM J. Optim.*, 20: 3384–3407.

10. Jeyakumar, V and Li, G. 2010. Characterizing robust set containments and solutions of uncertain linear programs without qualifications. *Oper. Res. Lett.*, 38: 188–194.

11. Jeyakumar, V, Li, GY and Lee, GM. 2011. A robust von Neumann minimax theorem for zero-sum games under bounded payoff uncertainty. *Oper. Res. Lett.*, 39: 109–114.

12. Jeyakumar, V, Ormerod, J and Womersley, RS. 2006. Knowledge-based semidefinite linear programming classifiers. *Optim. Methods Softw.*, 21: 693–706.

13. Khemchandani, R, Jayadeva and Chandra, S. 2009. Knowledge based proximal support vector machines. *Eur. J. Oper. Res.*, 195: 914–923.

14. Kumar, MA, Khemchandani, R, Gopal, M and Chandra, S. 2010. Knowledge based least squares twin support vector machines.*Inform. Sci.*, 180: 4606–4618.

15. Li, GY, Jeyakumar, V and Lee, GM. 2011. Robust conjugate duality for convex optimization under uncertainty with application to data classification. *Nonlinear Anal.*, 74: 2327–2341.

16. Mangasarian, OL. 2004/05. Knowledge-based linear programming. *SIAM J. Optim.*, 15: 375–382.

17. Mangasarian, OL and Musicant, RD. 2001. Lagrangian support vector machines. *J. Machine Learn. Res.*, 1: 161–177.

18. Rockafellar, RT. 1970. *Convex Analysis*, Princeton: Princeton University Press.

19. Schölkopf, B, Simard, PY, Smola, AJ and Vapnik, V. 1998. “Prior knowledge in support vector kernels”. In *Advances in Neural Information Processing Systems*, Edited by: Jordan, MI, Kearns, MJ and Solla, SA. 640–646. Cambridge: Vol. 10, MIT Press.

20. Vapnik, VN. 2000. *The Nature of Statistical Learning Theory*, 2nd, Berlin: Springer-Verlag.

21. Wang, L, Zhu, J and Zou, H. 2006. The doubly regularized support vector machine. *Stat. Sin.*, 16: 589–615.

22. Wang, L, Zhu, J and Zou, H. 2008. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24: 412–419.