Enhancing Transparency of Black-box Soft-margin SVM by Integrating Data-based Prior Information

Shaohan Chen\textsuperscript{1}, Chuanhou Gao\textsuperscript{1}, and Ping Zhang\textsuperscript{1}

\textsuperscript{1}School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China.

\textbf{Keywords:} Soft-margin SVM, black-box, transparency, data-based prior information, blast furnace

\section{Abstract}

The lack of transparency often makes the black-box models difficult to be applied to many practical domains. For this reason, the current work, from the black-box model input port, proposes to incorporate data-based prior information into the black-box soft-margin SVM model to enhance its transparency. The concept and incorporation mechanism of data-based prior information are successively developed, based on which the transparent or partly transparent SVM optimization model is designed and then solved through handily rewriting the optimization problem as a nonlinear quadratic programming problem. An algorithm for mining data-based linear prior information from data set is also proposed, which generates a linear expression with respect to two appropriate inputs identified from all inputs of system. At last, the proposed transparency strategy is applied to eight benchmark examples and two real blast furnace examples for effectiveness exhibition.

\section{1 Introduction}

Development of black-box modeling techniques, like support vector machine (SVM), neural networks, etc., has shown rather rapid in the past decades (Yuan et al., 2016; Zhao et al., 2015; Wu et al., 2013). This sort of techniques, compared to white-box modeling methods (also called mechanism-based modeling or first-principles modeling), works without any need of knowing the internal structure or details on variables interaction in systems considered, so they are suited to describe extremely complex objectives, such as human brain (Khosrowabadi et al., 2014), black hole (Grumiller et al., 2012), integrated industrial processes (Gao et al., 2012) and so on. Essentially, black-box modeling is an input-output data-based approach, and the model precision mainly
depends on data quality, model structure and parameters identification algorithm. In order to develop high-precision black-box models, it always needs reliable and representative data, smart mathematical treatment and efficient identification algorithms. All of these are challenging the development of the black-box modeling techniques. Intuitively, it is not always a good strategy to further develop advanced mathematical methods for the improvement of the black-box models precision. Moreover, even if black-box models are accurate enough, a clear insight into the reasoning made by them is not available. Namely, there is a severe lack of comprehensibility and transparency on the operating principle of black-box models. The loss of transparency makes it impossible to explain the model outputs as comprehensive knowledge, and neither to improve the model performance using known knowledge about systems. Since explanation is one of the most important aspects that affects end users to accept models, the applications of black-box models in practical domains are restricted greatly. This is specially true for areas like credit risk analysis and medical diagnosis, where the definite causal for making a decision is necessary and desired. Note that black-box models typically provide more accurate predictive results than white-box models. This motivates us to conduct the investigation of developing ways to enhance the transparency of black-box models so that the improved versions have both of advantages of black-box models and white-box models, i.e., high precision and transparency.

There are two ways to make a black-box model more transparent. One way is from the input port of the black box to integrate prior information into the model (Lauer et al., 2008; Qu et al., 2011; Borges et al., 2011), the other way is from the output port to extract comprehensive rules from the model (Martens et al., 2009; Chorowski et al., 2011; Huynh et al., 2011). A schematic diagram to display these two transparency ways is given in Fig. 1. In this paper, we will focus on the first strategy while the rules extracting strategy may be found in the recent review given by Barakat et al. (Barakat et al., 2010). The main reason is that the prior information is crucial for building models of the problem at hand. Its importance may be seen from no free lunch theorem (Wolpert et al., 1997) that states all algorithms perform the same when averaged over different problems. Lauer et al. (2008) even pointed out that a model without prior knowledge is an ill-posed problem. The priors incorporation is thought as the unique means for a model to be extended into practice (Niyogi et al., 1998) in the case that the data size is finite. More importantly, the inclusion of priors into the model can add both of transparency and precision. Furthermore, unlike rule extraction only giving transparency in knowledge reasoning, the prior information may refer to any aspect about the problem, like the structure, parameters, etc., so it can provide transparency from every aspect (Hu et al., 2007).

Prior information is any known information on the problem investigated beforehand. The forms of prior information are very diverse, some about the model structure, some about the inherent constraint, while some about the data example. For the successful incorporation of prior information, it is necessary to sort out them right. Broadly speaking, prior information is usually classified into two main categories: knowledge about the estimated function, such as smoothness (Vapnik et al., 1998), symmetry (Chen et al., 2008), monotonicity (Abu-Mostafa, 1990; Daniels et al., 2010), boundary constraint/input domain (Lauer et al., 2008; Mangasarian et al., 2004, 2007, 2008), etc., and knowledge on the data (Lauer et al., 2008b), such as quality of the data,
including that if the data set is persistent exciting, if the data set contains visible outliers, and if the data set is imbalanced (mainly in the case of classification problem, a high proportion of samples belongs to the same class). According to categories, the prior information is embedded into the model using different modes. The smoothness and symmetry may be used to define the model structure (Poggio et al., 1990; Aguirre et al., 2004); the monotonicity and concavity may be converted to the derivatives information of the estimated function, and are further incorporated in the form of the inequality constraints (Lauer et al., 2008a) into the model; the boundary constraint/input domain is embedded as equality or inequality constraint (Mangasarian et al., 2004, 2007, 2008). For the information of non-persistent exciting data, Niyogi et. al. (2008) proposed to create virtual samples to enlarge the data set, while for the imbalanced data, the idea of weighting the samples may trade off the data difference to weaken imbalance (Cristianini et al., 2000). In fact, it is not often easy to recognize the categories or the incorporation methods of the prior information explicitly. For examples, the boundary constraint/input domain is related to both the estimated function and the data, and the data knowledge is incorporated through affecting the estimated function.

Another relatively distinct incorporation method may be based on the black-box models that need to be enhanced transparency. Typical black-box models of concern include neural network and SVM. These two kinds of models have different structure, which leads to the incorporation means of priors very different. The former has a hierarchical structure with the incorporation of priors to modify the weights, bias, and/or minimize back-propagation errors. Towell et al. (1994) mapped “domain theories” in the form of propositional logic into network, and got stronger generalization ability. Daniels and his coworkers showed universal approximation capabilities of partial monotone (Minin et al., 2010) and monotone neural network (Daniels et al., 2010; Velikova et al., 2006), where partial monotonicity or monotonicity is incorporated by structure. Dugas et al. (2009) also confirmed that the generalization performance of neural network can be improved if the functional knowledge, like convexity and monotonicity, is incorporated. Hu et. al. (2007) presented three embedding modes for neural network from “structural”, “algorithm” and “data”, and ranked them in a descending order with respect to transparency. Hu et al. (2016) developed a general framework for deep neural network to incorporate and automatically optimize vast amount of fuzzy knowledge. As for the SVM model, it is of a constraints-optimization structure, so
the prior information is usually incorporated either in the form of equality constraint (Poggio et al., 1990; Aguirre et al., 2004; Cristianini et al., 2000) or of inequality constraint (Lauer et al., 2008a; Mangasarian et al., 2004, 2007, 2008). More details may be found in the review paper (Lauer et al., 2008b).

This work continues to enrich the methods of integrating the prior information into the black-box model. We take the soft-margin SVM (Scholkopf et al., 2002; Xu et al., 2013) as an example of black-box models to be incorporated with the priors for transparency enhancement. The main contributions include the development of an optimization-based linear priors mining algorithm from data, and the construction plus solving of the partly transparent soft-margin SVM model. In the model construction, we fully consider the correctness of the mined linear priors, and design the transparent model as the balance of maximizing the margin and minimizing the errors of the priors. In the model solving process, we rewrite the transparent model to have the same structure as the pure black-box soft margin SVM so that the commercial SVM software package (Chang et al., 2011) can be used directly. The mined linear priors are embedding into the black-box model in the form of inequality constraints (Mangasarian et al., 2004, 2007, 2008), which add transparency of the black-box model by affecting the model structure and further the solving algorithm. Additionally, the model precision is also enhanced from incorporating the mined linear priors, since some “important” samplings are expected to be classify right. We use eight benchmark examples and two real blast furnace examples to evaluate the performance of the proposed transparency method. Here, the motivation of considering applications to the blast furnace system is that this complex process is in urgent need of control decision-making based on the transparency of the blast furnace black-box models (Gao et al., 2014).

As an important economic item to any country, the blast furnace has attracted much attention both in academia and in industry. Various white-box models (Nogami et al., 2005; Jindal et al., 2007) and black-box models (Bhattacharya, 2005; Gao et al., 2009; Nurkkala et al., 2011) were alternatively developed for describing this complex process. Especially, the black-box models have emerged largely due to their high accuracy and dynamic on-line service in the recent decades. However, the closed-loop control is still in its infancy (Saxen et al., 2013), and the manual control of experienced foremen are the main dependence today to maintain stability and to control the production quality. The transparency of the blast furnace black-box models can thus provide a feasible solution for control decision-making, but still not losing the black-box models’ advantages. To ensure the black-box soft-margin SVM model working more practically, the following assumptions are made on pursuing the current research:

A1. In all probability, there are non-separable or mislabeled samples when classification is executed on real-world data;

A2. The prior information acquired from real-world data allows to be not exact as the true one.

Note that these two assumptions are so general in practice that this work can serve for extensive applications. The rest of this paper is organized as follows. Section II introduces preliminary on SVM and its soft-margin version. Then, formulation of data-based prior information is proposed in Section III. Next, Section IV presents a transparency pattern of the soft-margin SVM through incorporating data-based prior information. This is followed by some numerical experiments exhibition in Section
V, including eight benchmark examples and two real blast furnace examples. Finally, Section VI concludes this paper.

Throughout the paper, the bold typeface denotes a vector or a matrix while the normal typeface stands for a scalar. The transpose of a vector or a matrix is denoted by the superscript “\( \top \)”, while the superscript apostrophe denotes the derivative of a function with respect to its argument. For any two vectors \( \mathbf{z}_a = (z_{a1}, \cdots, z_{an})^\top \) and \( \mathbf{z}_b = (z_{b1}, \cdots, z_{bn})^\top \in \mathbb{R}^n \) the inequality \( \mathbf{z}_a \geq \mathbf{z}_b \) means \( z_{ai} \geq z_{bi}, \forall i \). More notations may be found in Table 1.

# 2 SVM and Its Soft-margin Version

## 2.1 Support Vector Machine

SVM is a kind of kernel-based black-box modeling method, the main idea of which is to construct a hyperplane in an imaginary high-dimensional feature space that could separate two different classes (labeled by the output \( y = +1 \) and \( y = -1 \), respectively) furthest (Vapnik et al., 1998). Mathematically, SVM works through implicitly defining a high-dimensional feature project \( \Phi : \mathbb{R}^n \rightarrow \mathcal{F} \) that maps the input pattern \( \mathbf{x} \in \mathbb{R}^n \) into the high-dimensional feature space \( \mathcal{F} (\dim(\mathcal{F}) \gg n) \), and then aiming to solve the optimization problem

\[
\max_{w, b} \frac{2}{\|w\|} \\
\text{s.t. } y_i(w^\top \Phi(x_i) + b) \geq 1, i = 1, \cdots, N. \tag{1}
\]

Here, \( w \) is the normal vector in \( \mathcal{F} \), \( b \in \mathbb{R} \) is the offset, and \( \frac{2}{\|w\|} \) measures the margin of classification. If \( w^\top \Phi(x_i) + b \geq +1 \) for a \( x_i \), then this \( x_i \) belongs to the class of \( y_i = +1 \) while if \( w^\top \Phi(x_i) + b \leq -1 \) then \( x_i \) belongs to the class of \( y_i = -1 \). The optimal solution of Eq. (1) renders the decision function to be

\[
f(x) = \text{sign}(w^\top \Phi(x) + b). \tag{2}
\]

Often, solving the optimization problem of Eq. (1) is realized by transforming it into a quadratic programming problem (Cristianini et al., 2000)

\[
\min_{w, b} \frac{1}{2} w^\top w \\
\text{s.t. } y_i(w^\top \Phi(x_i) + b) \geq 1, i = 1, \cdots, N. \tag{3}
\]

By introducing Lagrangian multipliers \( \alpha = (\alpha_1, \cdots, \alpha_N)^\top \in \mathbb{R}_+^N \), Eq. (3) can be rewritten as a saddle point problem

\[
\min_{w, b} \max_{\alpha} \frac{1}{2} w^\top w - \sum_{i=1}^N \alpha_i[y_i(w^\top \Phi(x_i) + b) - 1], \\
\text{s.t. } \alpha \geq 0_N. \tag{4}
\]
Table 1: Notations

| Symbol | Meaning |
|--------|---------|
| $F_1, F_2, F_3$ | Three functions related to the dual target margin, “positive and negative class” prior information, respectively |
| $\tilde{F}_2, \tilde{F}_3$ | Improved version of $F_2$ and $F_3$ through the Lagrangian multiplier $\alpha$ |
| $G_1, G_2$ | Two functions representing the terms related to the kernel function and the Lagrangian multipliers, respectively |
| $k(\cdot, \cdot)$ | $N$-dimensional kernel function vector |
| $l_1, l_2, l_3$ | The loss functions for incorrect classification, incorrect “positive class” and “negative class” prior information, respectively |
| $N$ | The number of training samples |
| $n$ | The dimension of data set |
| $\mathbb{R}^n$ | The space of $n$-dimensional real vectors |
| $\mathbb{R}^+_n$ | The set of $n$-dimensional real vectors consisting of all nonnegative entries |
| $\mathbb{R}^+_n$ | The set of $n$-dimensional positive real vectors |
| $g^+, g^-$ | Two maps from $\Gamma^+$ and $\Gamma^-$ to $\mathbb{R}^+_r$ and $\mathbb{R}^-_r$, respectively, expressing “positive and negative class” prior information |
| $r^+, r^-$ | Two positive integers |
| $t^+, t^-$ | The number of training samples contained in $\Gamma^+$ and $\Gamma^-$, respectively |
| $x^-, x^-$ | Sampling input to render “positive/negative class” prior information |
| $\alpha, \beta, \gamma$ | Lagrangian multiplier vectors |
| $\tilde{\beta}, \tilde{\gamma}$ | Improved version of Lagrangian multiplier vectors of $\beta$ and $\gamma$, respectively |
| $\Gamma^+, \Gamma^-$ | The subset of $\mathbb{R}^n$ containing sampling input $x^-, x^-$, respectively |
| $\mathcal{L}(\alpha)$ | The dual target margin |
| $\mathcal{L}_{ij}^+, \mathcal{L}_{ij}^-$ | Labeling linear “positive and negative class” prior information in the plane $X^{(i)}O X^{(j)}$ composed of two features $x^{(i)}$ and $x^{(j)}$ |
| $\zeta, \xi$ | Slack variable vectors to measure incorrectness of the mined “positive and negative class” prior information, respectively |
| $\theta$ | Kernel slack variable |
| $\lambda_1, \lambda_2, \lambda_3$ | Counterbalance constants |
| $v_+, v_-$ | $r^+$ and $r^-$ dimensional nonnegative real vectors, respectively |
| $\Phi$ | Feature map from $\mathbb{R}^n$ to high dimensional feature space |
| $\mathbf{0}_n, \mathbf{1}_n$ | $n$-dimensional vector with all entries equal to 0 and 1, respectively |
| s.t. | The abbreviation of “subject to” |
Further interchanging the order of optimization and also utilizing the “stationary” Karush-Kuhn-Tucker (KKT) conditions (Kuhn et al., 1951), the saddle point problem of Eq. (4) is equivalently written as its dual form

\[
\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \Phi(x_i)^\top \Phi(x_j),
\]

s.t. \( \alpha^\top y = 0, \, \alpha \geq 0_N \), (5)

where \( y = (y_1, \cdots, y_N)^\top \) represents the output vector. To avoid the dimensionality disaster, the inner product of high-dimensional vectors \( \Phi(x_i) \) and \( \Phi(x_j) \) is replaced by the “well-known” kernel function (Cristianini et al., 2000)

\[
k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j),
\]

the optimization problem thus changes to be

\[
\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j),
\]

s.t. \( \alpha^\top y = 0, \, \alpha \geq 0_N \), (7)

and the decision function follows

\[
f(x) = \text{sign} \left( \sum_{\alpha_i > 0} \alpha_i y_i k(x, x_i) + b \right).
\]

An available kernel function should satisfy Mercer’s conditions (Scholkopf et al., 2002) that the kernel matrix with the entry in the \( i \)th row and the \( j \)th column to be \( k(x_i, x_j) \) is positive semi-definite. Common ones include linear kernel, polynomial kernel, Gaussian radial basis kernel, sigmoid kernel, etc.

### 2.2 Soft-margin SVM

SVM in the form of Eqs. (7) and (8) works under the condition that the sample points should clearly fall into the area \( w^\top \Phi(x_i) + b \geq 1 \) or \( w^\top \Phi(x_i) + b \leq -1 \), but no points are allowed to fall between them. This kind of SVM is identified by hard-margin one. To handle mislabeled examples, the soft-margin version is proposed (Scholkopf et al., 2002) that introduces non-negative slack variables \( \xi_i \) (\( i = 1, \cdots, N \)) to measure the degree of misclassification of the data \( x_i \). The objective function then changes to evaluate the trade-off between a large margin and a small error penalty, i.e.,

\[
\frac{1}{2} w^\top w + C \sum_{i=1}^{N} \xi_i \quad (C \in \mathbb{R}^+ \text{ is the penalty factor}),
\]

and the constraint changes to be

\[
y_i (w^\top \Phi(x_i) + b) \geq 1 - \xi_i.
\]

A more flexible disposal route is to design moving hyperplanes, called \( \nu \)-SVM that has the optimization problem

\[
\min_{w, b, \rho \geq 0, \xi_i \geq 0} -\nu \rho + \frac{1}{2} w^\top w + \frac{1}{N} \sum_{i=1}^{N} \xi_i
\]

s.t. \( y_i (w^\top \Phi(x_i) + b) \geq \rho - \xi_i, \forall i \), (9)
where \( \nu \in [0, 1] \) represents the upper bound on the fraction of training errors or the lower bound on the fraction of support vectors. In the same way, by means of the Lagrangian multipliers, the final dual form reduces to be

\[
\max_{\alpha} \mathcal{L}(\alpha),
\]

\[
s.t. \quad \alpha^\top y = 0, \ 0_N \leq \alpha \leq \frac{1}{N} 1_N, \ \alpha^\top 1_N \geq \nu,
\]

(10)

where \( \mathcal{L}(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \). The decision function for \( \nu \)-SVM takes the same form of Eq. (8).

The optimization problem of Eq. (10) may be further relaxed by introducing the nonnegative kernel slack variable \( \theta \) (Xu et al., 2013), defined by the difference of the target margin \( \tau \) and the above dual target \( \mathcal{L}(\alpha) \), i.e.,

\[
\theta = \tau - \mathcal{L}(\alpha).
\]

(11)

Making a balance between the maximum margin and the minimum error penalty can create an new optimization problem

\[
\min_{\tau, \alpha \in \mathcal{A}, \theta \geq 0} -\tau + \lambda_1 \ell_1(\theta),
\]

\[
s.t. \quad \mathcal{L}(\alpha) \geq \tau - \theta,
\]

(12)

where \( \mathcal{A} = \{ \alpha | \alpha^\top y = 0, \ 0_N \leq \alpha \leq \frac{1}{N} 1_N, \ \alpha^\top 1_N \geq \nu \} \), the parameter \( \lambda_1 \in \mathbb{R}_+ \) acts as a counterbalance, and \( \ell_1(\theta) \) represents any loss function. The underlying SVM induced by Eq. (12) is referred to as soft-margin SVM in the following.

**Proposition 1.** Soft-margin SVM of Eq. (12) has the same solutions as \( \nu \)-SVM of Eq. (10).

**Proof.** This result is a special case of Proposition 2 in the paper (Xu et al., 2013). See the detailed proof there. \( \square \)

### 3 Data-based Prior Information

#### 3.1 Prior Information

Prior information incorporated into black-box models will add a high degree of transparency. Moreover, if the size of data is limited, this incorporation is thought as the sole means to improve the generalization performance of black-box models (Niyogi et al., 1998). Here, prior information is defined as follows.

**Definition 1.** (Qu et al., 2011). Prior information refers to any known information about or related to the concerning objects, such as data, knowledge, specifications, etc.

In this work, attention is mainly focused on the prior information related to data collected. However, it is not statistics of data, or characteristics that can be directly observed from data, such as imbalance, but logical implications acquired by data-mining
Consider a binary classification problem. Let the training examples be \( \{ (x_i, y_i) \}_{i=1}^{N} \), \( \Gamma^+ \subseteq \mathbb{R}^n \) and \( g^+: \Gamma^+ \rightarrow \mathbb{R}^{r^+} \), \( g^-: \Gamma^- \rightarrow \mathbb{R}^{r^-} \), where \( r^+ \) and \( r^- \) are two positive integers, then the nonlinear prior information is expressed as

\[
g^+(x) \leq 0_{r^+} \implies \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \geq b^*, \ \forall x \in \Gamma^+
\]  

(13)

to identify positive class points \( y = +1 \), and

\[
g^-(x) \leq 0_{r^-} \implies \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \leq -b^*, \ \forall x \in \Gamma^-
\]  

(14)

to classify negative class points \( y = -1 \). Here, \( b^* \in \bar{\mathbb{R}}^+ \) is often set 0 or 1 in practice. Mathematically, the above prior information, as an example of the “positive class” prior information of Eq. (13), means that

\[
\begin{align*}
  z^T k(x^+) & \leq 0_{r^+}, \\
  z^T k(x^+) + b & < b^*
\end{align*}
\]  

(15)

where \( x^+ \in \Gamma^+ \), \( k(x^+) = (k(x^+, x_1), \ldots, k(x^+, x_N))^T \) and \( z = (\alpha_1 y_1, \ldots, \alpha_N y_N)^T \). Moreover, if \( \Gamma^+ \) is a convex subset of \( \mathbb{R}^n \), and \( g^+(x^+) \) and \( k(x^+) \) are convex on \( \Gamma^+ \), the following result holds.

**Lemma 1.** (Mangasarian et al., 2008). The prior information expressed as Eq. (13) or Eq. (15) is equivalent to the result that there exists \( v_+ \in \mathbb{R}^{r^+} \), \( v_+ \geq 0_{r^+} \), such that

\[
\begin{align*}
  z^T k(x^+) + b - b^* + v_+^T g^+(x^+) & \geq 0.
\end{align*}
\]  

(16)

For the “negative class” prior information of Eq. (14), there is also the corresponding parallel result.

**Lemma 2.** (Mangasarian et al., 2008). The prior information expressed as Eq. (14) is equivalent to the result that there exists \( v_- \in \mathbb{R}^{r^-} \), \( v_- \geq 0_{r^-} \), such that

\[
\begin{align*}
  -z^T k(x^-) - b - b^* + v_-^T g^-(x^-) & \geq 0.
\end{align*}
\]  

(17)

The work of Mangasarian et al. (Mangasarian et al., 2004, 2007, 2008) successfully converted the prior information in the form of logical implication into the corresponding linear inequality constraint. Obviously, the latter is more easily to be incorporated into black-box models. In practical operation, it is possible that the prior information, expressed by Eq. (16) or (17), is not right. Hence, some nonnegative slack error variables \( \zeta \) and \( \varsigma \) are further introduced to relax them, respectively, which creates the final expressions in discrete time form as follows

\[
\begin{align*}
  z^T k(x^+_j) + b - b^* + v_+^T g^+(x^+_j) + \zeta_j & \geq 0, \ j = 1, \ldots, t^+
\end{align*}
\]  

(18)
and
\[-z^\top k(x_n^0) - b - b^* + v^\top g^-(x_n^-) + \varsigma_h \geq 0, \ h = 1, \ldots, t^-, \] (19)
where \(t^+ \leq N\) and \(t^- \leq N\). By taking these two inequalities as additional constraints, and also adding the linear penalties on the slack variables \(\varsigma\) and \(\varsigma\) to the objective function in model, incorporation of prior information into model will be finished. It has been shown in some numerical experiments that this incorporation can improve greatly precision of SVM models (Mangasarian et al., 2004, 2007, 2008).

### 3.2 Concept of Data-based Prior Information

To incorporate prior information into black-box models, it is necessary to acquire it firstly. A feasible solution to tackle this issue is to mine it from data. As said in the assumption A2, there is a deviation between the mined prior information and the true one in all probability. The acquisition of prior information is thus modeled by minimizing the loss of its incorrectness.

Consider any binary classification problem in which input-output pairs are \(\{(x_i, y_i)\}_{i=1}^N\) with \(x_i \in \mathbb{R}^n\) and \(y_i \in \{+1, -1\}\), and the decision function takes
\[y = \begin{cases} +1, & f(x) \geq b^*, \\ -1, & f(x) \leq -b^*, \end{cases} \] (20)
with \(b^* \geq 0\). Let \(g^+(x) \leq 0_{r^+}\) and \(g^-(x) \leq 0_{r^-}\) be the positive and negative class functions that need to be mined, respectively, then utilizing Eqs. (18) and (19) (Mangasarian et al., 2008) we can define two slack variable vectors, \(\varsigma\) and \(\varsigma\), to measure incorrectness of the mined prior information
\[\varsigma_j = b^* - [f(x_j^+) + v^\top g^+(x_j^+)], \ j = 1, \ldots, t^+ \] (21)
and
\[\varsigma_h = b^* - [v^\top g^-(x_h^-) - f(x_h^-)], \ h = 1, \ldots, t^- \] (22)
The loss induced by these two error variables may be evaluated by any loss function, similar to penalizing \(\theta\) in Eq. (12) for soft-margin SVM (Xu et al., 2013). Denote the loss function by \(l_2(\cdot)\) and \(l_3(\cdot)\) for \(\varsigma\) and \(\varsigma\), respectively, then the losses induced by them are \(\sum_{j=1}^{t^+} l_2(\varsigma_j)\) and \(\sum_{h=1}^{t^-} l_3(\varsigma_h)\). The penalty to the incorrectness may create the following optimization problem
\[
\min_{\varsigma \geq 0_{r^+}, \varsigma \geq 0_{r^-}} \lambda_2 \sum_{j=1}^{t^+} l_2(\varsigma_j) + \lambda_3 \sum_{h=1}^{t^-} l_3(\varsigma_h),
\] (23)
s.t. \(f(x_j^+) - b^* + v^\top g^+(x_j^+) + \varsigma_j \geq 0, \ j = 1, \ldots, t^+\), \(f(x_h^-) + b^* - v^\top g^-(x_h^-) - \varsigma_h \leq 0, \ h = 1, \ldots, t^-\),
where \(\lambda_2, \lambda_3 \in \mathbb{R}_+\) are constants, called counterbalance, like \(\lambda_1\) in Eq. (12). The solution may suggest two pieces of prior information: 1) If \(g^+(x) \leq 0_{r^+}\), then \(x\) has the class label \(y = +1\); 2) If \(g^-(x) \leq 0_{r^-}\), then \(x\) has the class label \(y = -1\). Thus, we have the following result.
Proposition 2. For any system with input-output pairs \( \{(x_i, y_i)\}_{i=1}^{N} \), where \( x_i \in \mathbb{R}^n \) and \( y_i \in \{+1, -1\} \), assume \( f(x) \) defined by Eq. (20) to be a decision function for addressing binary classification problem of this system. Then the prior information: 1) \( g^+(x) \leq 0 \Rightarrow y = +1 \) and 2) \( g^-(x) \leq 0 \Rightarrow y = -1 \) can be equivalently expressed as Eq. (23).

Proof. The result is straightforward from Lemma 1 and 2. \( \square \)

The prior information in the form of logical implication is converted into an optimization problem, for which the incorrectness resulting from data mining techniques is fully considered, so we define it by data-based prior information.

Definition 2. If prior information in the form of logical implication is acquired by any data-mining technique, and the incorrectness is minimized through Eq. (23), then the acquired prior information is called data-based prior information.

Remark 1. The validity of data-based prior information depends strongly on the decision function \( f(x) \) used, which may be any black-box model, but not limited to SVM. Although Proposition 2 assumes that \( f(x) \) is given, the decision function and data-based prior information may be optimized synchronously through Eq. (23) as long as the objective function and constraints are corrected and added upon requirement, respectively.

It should be noted that if the data-based priors deviate from the true ones far, then the synchronous optimization on the objective function of black-box model and the incorrectness of data-based priors will destroy the precision of the black-box model. At this point, it is difficult to tune the regularization parameters in the integrated model to enhance both of the transparency and precision. This may be also suggested by allowing non-separable or mislabelled samples and non-true priors in assumptions A1 and A2, respectively. However, from the viewpoint of advancing practical applications, the transparency enhancement is a little more urgent than high precision for black-box models, so the synchronous optimization is still a good strategy even if it may lead to slight precision reduction. This also constitutes our main motivation to integrate data-based priors into the black-box model to enhance transparency in the current work. Naturally, if the data-based priors are true, the synchronous optimization strategy is potential to result in the enhancement of both of transparency and precision of black-box models. We will revisit this point in the next section.

3.3 Algorithm for Mining Data-based Linear Prior Information

Generally, it is not easy to mine data-based prior information, especially when the feature variables relation contained in \( g^+(x) \) or \( g^-(x) \) is nonlinear or the input dimension \( n \) is high, even if at \( n = 3 \). For this reason, we only consider linear prior information generated from two feature variables of system. The main thought of mining data-based linear prior information is to map the sampling points in \( \mathbb{R}^n \) to a two dimensional plane \( X^{(i)}OX^{(j)} \), in which a linear relation between feature variables \( x^{(i)} \) and \( x^{(j)} \) is found to be able to separate a class of samplings completely while separate another class of samplings as many as possible. For example, let \( Z_{ij}^+ \) be a straight line representing the
Algorithm 1. Mining Data-based Linear Prior Information

Input: \( \{(x_k, y_k)\}_{k=1}^{N}, \mathbf{x} = (x^{(1)}, \ldots, x^{(n)}) \in \mathbb{R}^n, y_k \in \{+1, -1\} \)

Output: Data-based prior information

1: Data normalization;
2: Set \( \Gamma^+ = \{x_k | y_k = +1, \forall k\} \) and \( \Gamma^- = \{x_k | y_k = -1, \forall k\} \);
3: Find positive class prior information:
   for \( i = 1 \) to \( n - 1 \) : 1 do
      for \( j = i + 1 \) to \( n \) : 1 do
         \( \circ \) Map all \( x_k \) (\( k = 1, \cdots, N \)) to the plane \( X^{(i)}OX^{(j)} \) with images denoted by \( \text{Img}_{ij}x_k \);
         \( \circ \) Solve the following optimization problem
         \[
         \begin{align*}
         \max_{\phi, c} \quad & \text{Card} \left( \{(x^{(i)} + x^{(j)}) | x^{(i)} \cos \phi + x^{(j)} \sin \phi + c \leq 0\} \right) \\
         \text{s.t.} \quad & x^{(i)} \cos \phi + x^{(j)} \sin \phi + c > 0, \ \phi \in [0, 2\pi], \ c \in \mathbb{R}, \\
         & (x^{(i)}, x^{(j)}) \in \text{Img}_{ij} \Gamma^+, \ (x^{(i)}, x^{(j)}) \in \text{Img}_{ij} \Gamma^-.
         \end{align*}
         \]
         where \( \text{Card}() \) represents the number of elements in set;
         \( \circ \) Store the optimal results in \( \Omega_{(i,j; \hat{\phi}_{ij}, \hat{c}_{ij})} \);
      end for
   end for
4: Denote \( (i, j; \hat{\phi}_{ij}, \hat{c}_{ij}) = \arg \max_{i,j} \Omega_{(i,j; \hat{\phi}_{ij}, \hat{c}_{ij})} \);
5: Output positive class prior information
   \[
   \{k | x_k^{(i)} \cos \hat{\phi}_{ij} + x_k^{(j)} \sin \hat{\phi}_{ij} + \hat{c}_{ij} \leq 0 \implies y_k = +1\};
   \]
6: The same way produces negative class prior information.

positive class prior information function \( g^+(x) = 0_{r+} \), then it requires that all negative class samplings fall above \( L^+_{ij} \) while positive class samplings fall below \( L^+_{ij} \) as many as possible. The same principle could generate negative class prior information \( L^-_{ij} \). We formulate the mining process as Algorithm 1.

A look at Algorithm 1 might reveal that it needs to solve many optimization problems when the dimension of data, i.e., \( n \), is large. The scalability will thus become poor for high-dimensional data set. Actually, there are \( \frac{n(n-1)}{2} \) iterations in the Step 3,
so the computational complexity of the algorithm is $O(N \times \frac{n(n-1)}{2})$. Although it looks increasing promptly as $n$ becomes large, compared with the computational complexity of the SVM model, which is $O(n \times N^2)$ (Chang et al., 2011), the current one is rather small. The main reason is that the size of the data set is usually far larger than the dimension of the data set, i.e., $N \gg n$. Therefore, it will not lead to scalability disaster when the knowledge mining and SVM model solving are performed simultaneously using LIBSVM. We exhibit this algorithm through the following example.

**Example 1.** The Algorithm 1 is applied to Liver disorders, a public data set, to exhibit effect, in which there are 6 feature variables $x^{(1)}, \cdots, x^{(6)}$, and 345 recordings with 200 positive class points while 145 negative class points. We select 70% sampling points randomly and feed them into the mining algorithm. Table 2 reports the mined positive class linear prior information along every projected two dimensional plane. It is clear that the straight line

$$L^+_{(1,4;\hat{\phi}_{14},\hat{c}_{14})} : 0.6347x^{(1)} - 0.7728x^{(4)} - 0.0156 = 0$$

in the plane $X^{(1)}OX^{(4)}$ is the expected one. In the same way, we can obtain negative class prior information in the plane $X^{(3)}OX^{(5)}$ to be

$$L^-_{(3,5;\hat{\phi}_{35},\hat{c}_{35})} : -0.1288x^{(3)} + 0.9917x^{(5)} + 0.0025 = 0.$$

**Remark 2.** The linear prior information actually represents a boundary in a plane related to two feature variables of system. Every boundary serves for the largest degree of separation between two classes in the plane constructed by two particular variables. For “positive”/“negative” boundary, all negative/positive class samplings fall above it while positive/negative class samplings fall below it as many as possible.

**Remark 3.** It is possible to generate multi sets of positive priors or of negative ones for an objective system. As an example of positive prior information, this takes place when there are more than two projected two dimensional planes in which the positive “boundaries” frame the same number positive samplings below them. In this case, the multi sets of priors can be applied at the same time without worrying that they are conflicting since they come from different planes.

**Remark 4.** Compared with the high dimensional nonlinear model of system, the linear prior information only including two feature inputs is quite simple. However, it provides an available way to mine the knowledge hidden in data. An interesting and challenging issue for future research is to include more inputs or to use nonlinear function with respect to inputs to model prior information. It may be relatively simple to construct a separating plane as prior information that is captured by a linear function with respect to three feature inputs.

---

1. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets
| \(X^{(i)}OX^{(j)}\) | \(\mathcal{L}^{+}_{(i,j;\hat{\phi}_{ij},\hat{c}_{ij})}\) | \(\Omega_{(i,j;\hat{\phi}_{ij},\hat{c}_{ij})}\) |
|----------------|-----------------|-----------------|
| \(X^{(1)}OX^{(2)}\) | \(-0.9422x^{(1)} + 0.3350x^{(2)} + 0.7347 \leq 0\) | 5 |
| \(X^{(1)}OX^{(3)}\) | \(0.1700x^{(1)} + 0.9854x^{(3)} - 0.1651 \leq 0\) | 10 |
| \(X^{(1)}OX^{(4)}\) | \(0.6347x^{(1)} - 0.7728x^{(4)} - 0.0156 \leq 0\) | 13 |
| \(X^{(1)}OX^{(5)}\) | \(0.5403x^{(1)} + 0.8415x^{(5)} - 0.2648 \leq 0\) | 4 |
| \(X^{(1)}OX^{(6)}\) | \(-0.9900x^{(1)} + 0.1411x^{(6)} + 0.8689 \leq 0\) | 6 |
| \(X^{(2)}OX^{(3)}\) | \(-0.7259x^{(2)} - 0.6878x^{(3)} + 0.7686 \leq 0\) | 5 |
| \(X^{(2)}OX^{(4)}\) | \(-0.7910x^{(2)} - 0.6119x^{(4)} + 0.8974 \leq 0\) | 5 |
| \(X^{(2)}OX^{(5)}\) | \(-0.8968x^{(2)} - 0.4425x^{(5)} + 0.8473 \leq 0\) | 3 |
| \(X^{(2)}OX^{(6)}\) | \(-0.8011x^{(2)} + 0.5985x^{(6)} + 0.6779 \leq 0\) | 2 |
| \(X^{(3)}OX^{(4)}\) | \(0.9950x^{(3)} + 0.0998x^{(4)} - 0.0634 \leq 0\) | 6 |
| \(X^{(3)}OX^{(5)}\) | \(0.9801x^{(3)} + 0.1987x^{(5)} - 0.0510 \leq 0\) | 4 |
| \(X^{(3)}OX^{(6)}\) | \(-0.5048x^{(3)} + 0.8632x^{(6)} + 0.1437 \leq 0\) | 6 |
| \(X^{(4)}OX^{(5)}\) | \(-0.8011x^{(4)} + 0.5985x^{(5)} + 0.3473 \leq 0\) | 7 |
| \(X^{(4)}OX^{(6)}\) | \(-0.5885x^{(4)} + 0.8085x^{(6)} + 0.2136 \leq 0\) | 9 |
| \(X^{(5)}OX^{(6)}\) | \(-0.9422x^{(5)} + 0.3350x^{(6)} + 0.4111 \leq 0\) | 5 |
4 Incorporation of Data-based Prior Information into Soft-margin SVM

In this section, data-based prior information is incorporated into soft-margin SVM for the purposes of increasing its transparency and also maintaining high precision.

For soft-margin SVM model of Eq. (12), it requests to minimize the margin error, while for data-based prior information model of Eq. (23) it requests to minimize the priors incorrectness. Their incorporation is naturally made through minimizing the sum of two objective functions, i.e., making a trade-off between a large margin and small error penalties, which creates the following optimization problem

$$\min_{\tau, \alpha \in \mathcal{A}, \theta \geq 0, \zeta \geq 0, \varsigma \geq 0} -\tau + \lambda^\top \ell,$$

s.t.

$$\mathcal{L}(\alpha) \geq \tau - \theta,$$

$$f(x_j^+) - b^* + v_+^\top g^+(x_j^+) + \zeta_j \geq 0, j = 1, \cdots, t^+,$$

$$f(x_h^-) + b^* - v_-^\top g^-(x_h^-) - \varsigma_h \leq 0, h = 1, \cdots, t^-,$$

where \(\lambda = (\lambda_1, \lambda_2, \lambda_3)^\top, f(\cdot) = z^\top k(\cdot) + b \) (See Eq. (15) for the definitions of \(z\) and \(k(\cdot)\)) and

$$\ell = \left(l_1(\theta), \sum_{j=1}^{t^+} l_2(\zeta_j), \sum_{h=1}^{t^-} l_3(\varsigma_h)\right)^\top.$$

This model will act as the benchmark model that can enhance the transparency but without loss of precision of the black-box soft-margin SVM. We refer to it and its equivalent models as partly transparent soft-margin SVM, abbreviated to pTsm-SVM.

A common strategy to solve the above optimization problem is to convert it into the corresponding dual form, then we have the following proposition.

**Proposition 3.** The solution of pTsm-SVM in Eq. (24) is the same as that of the following optimization problem

$$\min_{\alpha} \max_{\beta, \gamma} F_1(\alpha) + F_2(\alpha, \beta) + F_3(\alpha, \gamma) + \lambda^\top \ell,$$

where

$$\left\{
\begin{array}{l}
F_1(\alpha) = -\mathcal{L}(\alpha), \\
F_2(\alpha, \beta) = -\sum_{j=1}^{t^+} \beta_j [f(x_j^+) - b^* + v_+^\top g^+(x_j^+)], \\
F_3(\alpha, \gamma) = \sum_{h=1}^{t^-} \gamma_h [f(x_h^-) + b^* - v_-^\top g^-(x_h^-)],
\end{array}\right.$$ \hspace{1cm} (26)

and

$$\ell = l - \left(l_1'(\theta)\theta, \sum_{j=1}^{t^+} l_2'(\zeta_j)\zeta_j, \sum_{h=1}^{t^-} l_3'(\varsigma_h)\varsigma_h\right)^\top$$

with \(l_1'(\cdot), l_2'(\cdot), l_3'(\cdot)\) representing the corresponding derivatives with respect to their respective argument. The decision variables \(\beta = (\beta_1, \cdots, \beta_{t^+})^\top \geq 0_{t^+}\) and \(\gamma = (\gamma_1, \cdots, \gamma_{t^-})^\top \geq 0_{t^-}\) are the Lagrangian multipliers.
Proof. The conversion from Eq. (24) to Eq. (25) is easily realized by constructing the Lagrangian function of the former, and then utilizing the corresponding KKT conditions, very similar to the conversion from Eq. (3) to Eq. (4).

Remark 5. The optimization model of Eq. (25) will degenerate to be

\[
\min_{\alpha} \max_{\beta, \gamma} F_1(\alpha) + F_2(\alpha, \beta) + F_3(\alpha, \gamma)
\]

(27)

if all the loss functions are selected as the hinge loss, i.e., \(l(\cdot) = \max(0, \cdot)\) (Xu et al., 2013). In this case, the decision variables satisfy \(0_{t^+} \leq \beta \leq \lambda_2 1_{t^+}, 0_{t^-} \leq \gamma \leq \lambda_3 1_{t^-}\) and \(\alpha \in \mathcal{A}\).

Remark 6. For the degenerated pTsm-SVM in Eq. (27), the objective consists of three functions, \(F_1(\alpha), F_2(\alpha, \beta)\) and \(F_3(\alpha, \gamma)\), which respectively measure the contributions of pure SVM, “positive class” and “negative class” prior information to the optimization objective. The first term represents the black-box part of model while the latter two terms express the white-box part of model. The ratio of “black to white” in the model can be controlled through setting feasible domains of the decision variables.

Remark 7. Besides enhancing transparency, the priors incorporation into black-box soft-margin SVM also have potential to improve the model performance if they are true. On the one hand, these additional extra constraints can reduce the feasible domain of soft-margin SVM greatly, which provides larger opportunity to find global solutions; on the other hand, the acquired priors mean that some “important” samplings are singled out from all training points, and these “important” samplings are expected to be classified right with the trained model. Hence, these samplings will have a stronger effect on identifying model parameters.

A further look at the expressions of \(F_1(\alpha), F_2(\alpha, \beta)\) and \(F_3(\alpha, \gamma)\) in Eq. (26) reveals that \(F_1(\alpha)\) has different structure regarding \(\alpha\) from \(F_2(\alpha, \beta)\) and \(F_3(\alpha, \gamma)\). This may lead to large difficulty in utilizing the commercial SVM software package, like LibSVM (Chang et al., 2011). To be applied conveniently, we reformulate the benchmark model of Eq. (24) for pTsm-SVM as follows.

**Proposition 4.** The optimization problem

\[
\min_{\tau, \alpha \in \mathcal{A}, \theta \geq 0, \xi \geq 0, \zeta \geq 0} -\tau + \lambda^T l,
\]

(28)

\[
\text{s.t. } \mathcal{L}(\alpha) \geq \tau - \theta,
\]

\[
\alpha^+_j y^+_j [f(x^+_j) - b^* + v^+_j g^+(x^+_j) + \zeta_j] \geq 0, \forall j,
\]

\[
\alpha^-_h y^-_h [f(x^-_h) - b^* - v^-_j g^-(x^-_h) - \zeta_h] \geq 0, \forall h
\]

is equivalent to pTsm-SVM in Eq. (24), where \(\alpha^+_j\) and \(\alpha^-_h\) are entries in \(\alpha\) corresponding to positive and negative class samplings, respectively.

**Proof.** Since \(\forall j, \alpha^+_j \geq 0\) and \(y^+_j = +1\), and \(\forall h, \alpha^-_h \geq 0\) and \(y^-_h = -1\), the last two constraints in this equation are essentially the same as the corresponding two constraints emerging in Eq. (24). Therefore, the result is true. \(\square\)
Consider all losses in Eq. (28) induced by the hinge loss function, then we have the proposition as follows.

**Proposition 5.** The following optimization problem

\[
\min_{\alpha \in A} \max_{\beta, \gamma} F_1(\alpha) + \tilde{F}_2(\alpha, \tilde{\beta}) + \tilde{F}_3(\alpha, \tilde{\gamma})
\]  

(29)

shares the same solution with \( pTsm\)-SVM in Eq. (28) if all the losses induced by the slack variables \( \theta \), \( \zeta \) and \( \varsigma \) obey the rule of the hinge loss, where

\[
\begin{align*}
\tilde{F}_2(\alpha, \tilde{\beta}) &= -\sum_{j=1}^{t^+} \tilde{\beta}_j \alpha_j^+ [f(x_j^+) - b^* + \nu^+_j g^+(x_j^+)], \\
\tilde{F}_3(\alpha, \tilde{\gamma}) &= \sum_{h=1}^{t^-} \tilde{\gamma}_h \alpha_h^- [f(x_h^-) + b^* - \nu^-_h g^-(x_h^-)],
\end{align*}
\]

\( \tilde{\beta} \) and \( \tilde{\gamma} \) are the corresponding Lagrangian multiplier vectors.

**Proof.** Eq. (29) is the dual form of Eq. (28) under the given conditions, and they thus have the same solution. \( \Box \)

**Remark 8.** The decision variables \( \tilde{\beta} \) and \( \tilde{\gamma} \) in Eq. (29) have bounds as \( \forall j, 0 \leq \tilde{\beta}_j \leq \frac{\lambda_j}{\alpha_j} \) and \( \forall h, 0 \leq \tilde{\gamma}_h \leq \frac{\lambda_h}{\alpha_h} \).

For the convenience of using the commercial software package, further rewriting the optimization objective of Eq. (29) yields

\[
\min_{\alpha \in A} \max_{\beta, \gamma} G_1(\alpha, \tilde{\beta}, \tilde{\gamma}) - G_2(\alpha, \tilde{\beta}, \tilde{\gamma}),
\]  

(30)

where

\[
\begin{align*}
G_1(\alpha, \tilde{\beta}, \tilde{\gamma}) &= F_1(\alpha) - \sum_{j=1}^{t^+} \tilde{\beta}_j \alpha_j^+ \mathbf{z}^\top \mathbf{k}(x_j^+) \\
&\quad + \sum_{h=1}^{t^-} \tilde{\gamma}_h \alpha_h^- \mathbf{z}^\top \mathbf{k}(x_h^-), \\
G_2(\alpha, \tilde{\beta}, \tilde{\gamma}) &= \sum_{j=1}^{t^+} \tilde{\beta}_j \alpha_j^+ [b - b^* + \nu^+_j g^+(x_j^+)] \\
&\quad - \sum_{h=1}^{t^-} \tilde{\gamma}_h \alpha_h^- [b + b^* - \nu^-_h g^-(x_h^-)].
\end{align*}
\]

Clearly, the degenerated \( pTsm\)-SVM in Eq. (30) has a very similar structure to the “pure” black-box SVM in Eq. (7) with \( G_1(\alpha, \beta, \gamma) \) and \( G_2(\alpha, \beta, \gamma) \) individually representing the terms related to the kernel function and to the Lagrangian multipliers. This makes Eq. (30) look like a “standard” SVM so that the corresponding nonlinear QP problem can be easily solved using LibSVM (Chang et al. 2011).

The incorporation mechanism of priors into the black-box soft-margin SVM shown above combines ideas from Mangasarian et al. (Mangasarian et al. 2004, 2007, 2008) to be solved. This means the proposed \( pTsm\)-SVM to be more suitable for capturing practical nonlinear problems. Moreover, the QP problem is further handled according to **Propositions 4** and **5**, and finally takes the form of Eq. (30) which has the same structure as a “pure” black-box SVM and
Table 3: Benchmark Data Sets Information

| Name of Data Set | Size of Recordings (positive/negative) | Number of Feature Variables |
|------------------|----------------------------------------|----------------------------|
| Australian       | 690 (307/383)                           | 14                         |
| Breast cancer    | 683 (239/444)                           | 10                         |
| Diabetes         | 768 (268/500)                           | 8                          |
| German           | 1000 (300/700)                          | 24                         |
| Heart            | 270 (120/150)                           | 13                         |
| Ionosphere       | 351 (225/126)                           | 34                         |
| Liver disorders  | 345 (200/145)                           | 6                          |
| Sonar            | 208 (97/111)                            | 60                         |

can thus be solved directly utilizing the existing software packages developed for the standard SVM. More importantly, the incorporation pattern encourages the SVM model to perform classification task following rules, i.e., the mined priors, so the black-box SVM model changes to be interpretable. As a result, the $pTsm$-SVM has the advantage of white-box models as well as of black-box models, i.e., transparency together with high precision.

5 Numerical Experiments

In this section, the degenerated $pTsm$-SVM in Eq. (30) is used to model some practical data sets, including 8 benchmark data sets (all available at the same website as given in Example 1) and 2 real blast furnace data sets. For these 8 benchmark examples, two-class classification problems are addressed while for 2 blast furnace examples, three-class classification issues are tackled. The priors for every data set are linear and acquired through the proposed mining algorithm in subsection 3.3. For the kernel function in Eq. (30), the Gaussian radial basis kernel defined by

$$k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2)$$  \hspace{1cm} (31)

is used. The parameters training is made through grid search together with five-fold crossing validation for the purpose of reducing over-fitting phenomenon.

5.1 Benchmark Examples

The benchmark examples include 8 public data sets: Australian, Breast cancer, Diabetes, German, Heart, Ionosphere, Liver disorders and Sonar. Some basic information about them, like size and number of feature variables, is exhibited in Table 3. For every data set, the recordings are segmented into two groups, one group including 70% samplings as training set, and the other group including the remaining 30% samplings as testing set. The training set serves for generating linear priors and learning parameters while the testing set works for evaluating performance of the degenerated $pTsm$-SVM in Eq. (30).
Table 4: Acquired Prior Information and Classification Results of Benchmark Data Sets

| Name of Data Set       | Linear Prior Information                                                                 | ATA$^\dagger$ (%) with Priors | ATA (%) without Priors | p-value of the t-test |
|-----------------------|------------------------------------------------------------------------------------------|--------------------------------|------------------------|-----------------------|
| Australian            | $g^+(x) = -0.3037x^{(5)} - 0.9516x^{(14)} + 0.3085 \leq 0 \implies y = +1$\n
  $g^-(x) = 0.7087x^{(8)} - 0.7055x^{(10)} + 0.0176 \leq 0 \implies y = -1$ | 85.22 ± 2.70                      | 82.54 ± 6.40                  | 0.03                  |
| Breast cancer         | $g^+(x) = 0.8855x^{(1)} - 0.4646x^{(9)} + 0.2890 \leq 0 \implies y = +1$\n
  $g^-(x) = 0.9211x^{(1)} + 0.3894x^{(8)} - 0.0746 \leq 0 \implies y = -1$ | 97.34 ± 0.53                      | 96.85 ± 0.58                  | 0.00                  |
| Diabetes              | $g^+(x) = -0.9900x^{(2)} + 0.1411x^{(7)} + 0.9467 \leq 0 \implies y = +1$\n
  $g^-(x) = 0.3624x^{(2)} + 0.9320x^{(8)} - 0.2004 \leq 0 \implies y = -1$ | 76.67 ± 1.66                      | 75.92 ± 1.26                  | 0.04                  |
| German                | $g^+(x) = 0.6347x^{(2)} - 0.7728x^{(4)} + 0.3679 \leq 0 \implies y = +1$\n
  $g^-(x) = 0.0875x^{(4)} - 0.9962x^{(17)} + 0.9812 \leq 0 \implies y = -1$ | 75.83 ± 2.44                      | 74.83 ± 2.23                  | 0.05                  |
| Heart                 | $g^+(x) = -0.3739x^{(10)} + 0.9275x^{(8)} - 0.0547 \leq 0 \implies y = +1$\n
  $g^-(x) = 0.7756x^{(5)} - 0.6317x^{(8)} + 0.4139 \leq 0 \implies y = -1$ | 84.25 ± 3.74                      | 83.50 ± 4.78                  | 0.26                  |
| Ionosphere            | $g^+(x) = -0.6663x^{(5)} + 0.7457x^{(27)} + 0.3465 \leq 0 \implies y = +1$\n
  $g^-(x) = x^{(5)} - 0.5210 \leq 0 \implies y = -1$ | 93.96 ± 2.23                      | 93.49 ± 1.69                  | 0.21                  |
| Liver disorders       | $g^+(x) = 0.6347x^{(1)} - 0.7728x^{(4)} - 0.0156 \leq 0 \implies y = +1$\n
  $g^-(x) = -0.1288x^{(3)} + 0.9917x^{(5)} + 0.0025 \leq 0 \implies y = -1$ | 73.10 ± 3.70                      | 70.80 ± 3.52                  | 0.02                  |
| Sonar                 | $g^+(x) = 0.9553x^{(13)} + 0.2955x^{(20)} - 0.2426 \leq 0 \implies y = +1$\n
  $g^-(x) = -0.5748x^{(11)} - 0.8183x^{(27)} + 0.9725 \leq 0 \implies y = -1$ | 85.72 ± 3.59                      | 83.02 ± 4.55                  | 0.04                  |

$^\dagger$ represents Average Testing Accuracy.
Table 5: Classification Results on the Benchmark Data Sets with Other Models

| Name of Data Set | degenerated pTsm-SVM | Soft-margin MKL (Xu et al., 2013)§ | ℓ₂trStMKL (Liu et al., 2013)† |
|------------------|----------------------|-----------------------------------|--------------------------------|
| Australian       | 85.22 ± 2.70         | 86.23 ± 1.94                      | ∅                             |
| Diabetes         | 76.67 ± 1.66         | 76.35 ± 2.79                      | ∅                             |
| German           | 75.83 ± 2.44         | ∅                                 | 74.40 ± 1.00                   |
| Heart            | 84.25 ± 3.74         | 81.60 ± 4.21                      | 85.10 ± 1.40                   |
| Ionosphere       | 93.96 ± 2.23         | 91.33 ± 2.82                      | 94.70 ± 1.10                   |
| Liver disorders  | 73.10 ± 3.70         | ∅                                 | 66.20 ± 2.40                   |
| Sonar            | 85.72 ± 3.59         | ∅                                 | 83.30 ± 2.60                   |

§ ATA over 10 times random experiments; † ATA over 30 times random experiments.

The experiments begin with normalizing all input features in the training recordings to the range [0, 1]. Then we apply the proposed mining algorithm to acquire the linear priors for every data set. Shown in Table 4 are the results. After substituting them into Eq. (30), respectively, there are (ν, σ, ˜β, ˜γ) left to be trained. Here, we use grid search to find these optimal parameters. For the first parameter ν, its physical meaning implies it should not be too high, i.e., requesting enough good training, but to avoid getting in over-fitting, it may not be too low, i.e., training not allowed to be too good. We thus set a 10-point uniform discretization in [ν_{min}, ν_{max}] as the searching range for finding it with ν_{min} = 0.1 and ν_{max} depending on the specific example, calculated by \( ν_{max} = \frac{2^{\min(N^+, N^-)}}{N} \) (Chen et al., 2005) where N^+ and N^- represent the number of positive samplings and negative ones, respectively. As for the other three parameters, the searching ranges are \{2^{-3}, 2^{-2}, \ldots, 2^{5}, 2^{6}\} for σ, and \( \frac{1}{N} \) times of some points in [0, 1] for the components of ˜β and ˜γ. We consult the penalty factor of slack variables ξ_i (i = 1, \ldots, N) in Eq. (9) to set the searching range for ˜β and ˜γ like so.

After finishing learning these parameters through grid search and five-fold cross-validation, we further use the testing samplings to evaluate the performance of the degenerated pTsm-SVM. To make the results more convincing, we have carried out 10 times random experiments for every example, and calculated their average testing accuracy values and the corresponding standard deviations, as shown in Table 4. Here, the testing accuracy is defined by the ratio of the number of testing samplings to be classified right to the total number of testing samples. In each experiment, 70% samplings are selected randomly as the training set while the remaining 30% samplings are set as the testing set. The average testing accuracy (ATA) reported in Table 4 suggests that the degenerated pTsm-SVM basically can perform the 2-class classification task for these 8 benchmark examples well, high ATA but low standard deviations.

To further exhibit the performance of the degenerated pTsm-SVM, we have made some parallel experiments on these data sets using the soft-margin SVM model, i.e., without priors incorporated. The testing results are also reported in Table 4. It is clear that the degenerated pTsm-SVM outperforms the soft-margin SVM for all benchmark data sets according to the ATA. Moreover, the stability of the ATA, characterized by the standard deviation, for some data sets is also strengthened after incorporating priors.
such as for Australian, Heart and Sonar data sets where much smaller standard deviations emerge. For Breast cancer data set, the ATA stability basically keeps unchanged; but for Diabetes, German, Ionosphere and Liver disorders data sets, the ATA stability changes a little weaker when priors are integrated. The conflict phenomena reflected by the ATA and standard deviation for the last four mentioned data sets mean it difficult to say the inclusion of linear priors playing a positive role on improving precision for them. Even though for the Australian-like data sets, it is still difficult to say that the degenerated $pTsm$-SVM must have higher precision than the soft-margin SVM, as the extreme case $(85.22 - 2.70)$ in the former model is apparently lower than the extreme case $(82.54 + 6.40)$ in the latter one. To achieve rigorous comparisons, we make a paired Student’s t-test on the classification results produced by the used two kinds of models. The test results (p-values) are also reported in Table 4, where the p-value represents the probability that the ATA of the soft-margin SVM model is no less than the ATA of the degenerated $pTsm$-SVM with the significant level of 0.05. Clearly, except the data sets of Heart and Ionosphere, the other 6 data sets exhibit that the integration of the mined priors can improve the precision of the soft-margin SVM model in very high probability, i.e., in these 6 data sets the degenerated $pTsm$-SVM is statistically significantly better than the soft-margin SVM. Therefore, from these 6 examples, it might suggest that the degenerated $pTsm$-SVM, on the one hand, has higher transparency than the soft-margin SVM (the structure and the solving algorithm of the black-box soft-margin SVM are changed due to incorporation of the mined linear priors, and moreover, the mined linear priors are highly related to the background of the corresponding data set); on the other hand, the former has larger possibility to make right classifications than the latter. As for the data sets of Heart and Ionosphere, although the degenerated $pTsm$-SVM is more transparent than the soft-margin SVM, it cannot outperform the latter statistically significantly in precision, the p-values only being 26% and 21%, respectively. The possible reason for these two high p-values may be that the linear priors mined for these two data sets are not so good, even not true, which sometimes play a constructive role while sometimes play a negative role. A solution to overcome this issue may be either to integrate other priors with respect to more features instead or to relax the current ones so that a little less positive/negative points are included below the “positive”/“negative” boundaries. Despite the possibility of slight loss in precision for these two data sets, the degenerated $pTsm$-SVM is still a good alternative due to its transparency and larger potential of practical applications.

To exhibit the reliability of the degenerated $pTsm$-SVM in precision, we present the experimental results on some of those benchmark examples produced by competing state-of-the-arts in Table 5. The model in [Xu et al., 2013] is a soft-margin multiple kernels SVM while the model in [Liu et al., 2013] is multiple kernels SVM integrating radius information. These two kinds of models were asserted to be able to outperform other similar models, like MKL, $\ell_p$MKL, etc. As can be seen from Table 5 for the degenerated $pTsm$-SVM and the listed two models each has its own merits, either in ATA or in standard deviation. However, it needs to mention that it is not quite fair to compare the current classification results with those generated by Soft-margin MKL or $\ell_2$trStMKL, since the testing samplings are not the same during every random experiment. We list them here not for solid comparisons but only for reference.

In summary, from the viewpoint of model precision, the degenerated $pTsm$-SVM
seems not significantly superior to some competing state-of-the-arts, and even a simple multi-layer neural network might produce better accuracy in some of the benchmark data sets. However, the largest advantage for the proposed model is that it has transparency while the others are “black”. In the current modeling framework, there are practical domains knowledge, i.e., the mined linear priors, for every data set before modeling them, which can be thought as known information about the modeling object. The incorporation of these priors into the black-box model, soft-margin SVM, has a large effect on the model structure and the solving algorithm, which results in the transparency enhancement of the black-box soft-margin SVM. Actually, it is not our original intension to expect the degenerated $pTsm$-SVM better than the existing best black-box model in precision when used for the benchmark data sets. The main contribution of this paper is to provide a way for adding transparency of black-box models, and the models performance comparison should be made between the transparent model and the corresponding black-box model. It is interesting to observe the effect of the mined linear priors, maybe in the form of logical implications as given in Eqs. (13) and (14), are incorporated into other black-box models, such as neural network, MKL, etc. To ensure the model performance better, it naturally needs to mine priors as accurate as possible from the data sets, which constitutes one of our main concerns in the future research. In addition, it is very time-consuming for the degenerated $pTsm$-SVM to be trained, including priors mining plus SVM training. As an example of the Heart data set, the average processing time is 15.1s for the degenerated $pTsm$-SVM while 3.9s for the soft-margin SVM. It is possible to avoid this point by achieving the optimal parameters related to the priors, like $\tilde{\beta}$ and $\tilde{\gamma}$ in Eq. (30), based on theoretical analysis. The effort towards this target is on the way.

5.2 Real Blast Furnace Examples

5.1 Blast furnace problem formulation

Blast furnace (BF) is a crucial operation unit in the integrated route of steel production, whose main function is to generate molten iron, often called pig iron or hot metal, through chemically reducing and physically converting iron oxides. The whole iron-making process involves various chemical reactions and transport phenomena, which together with hostile measurement environment make the blast furnace reactor too complex to be controlled effectively. As far as the blast furnace system is concerned, the control often means to control the hot metal temperature and compositions, like silicon content and sulfur content in hot metal, within acceptable bounds. Thereinto, the hot metal silicon content is the most concerning control object. On the one hand, it indicates the in-furnace thermal status since the silicon transfer from silica to the hot metal occurs as an endothermic reaction that could affect the bottom of the furnace (i.e., the hearth) and further influence the hot metal temperature; on the other hand, it reflects the consumption of coke, which says that an increasing silicon content often means a surplus of coke while a decreasing one implies a depletion of coke. From the viewpoint of energy saving, it is desired to operate the blast furnace at low hot metal silicon content, but still avoiding the risk of chilled hearth (Saxen et al., 2013), i.e., not too low. The silicon content appears so relevant to the hot metal quality and fuel consumption that
its control issue becomes a main concern in the blast furnace operations.

Let \( z \) represent the hot metal silicon content in the context. Its acceptable supremum and infimum are denoted by \( z_{\text{sup}} \) and \( z_{\text{inf}} \), respectively. The silicon control problem aims to manipulate the blast furnace inputs so that the hot metal silicon content fall into the range \( [z_{\text{inf}}, z_{\text{sup}}] \), and had better approach \( z_{\text{inf}} \), i.e., in \( \{z | z \in [z_{\text{inf}}, z_{\text{sup}}] \} \cap \{z | z - z_{\text{inf}} \ll z_{\text{sup}} - z \} \). Here, analogous to our earlier work (Gao et al., 2014), this control problem is transformed into a three-class classification task. The degenerated \( pTsm-SVM \) in Eq. (30) applied to the blast furnace system will generate some operation patterns to render the silicon content in the desired range. Mathematically, suppose the silicon content in \( [0, z_{\text{inf}}) \) as low silicon, in \( (z_{\text{inf}}, z_{\text{sup}}] \) as proper silicon, and in \( (z_{\text{sup}}, 1] \) as high silicon, then the operation patterns can be written as

\[
\text{Patterns: } \text{Inputs Set } X^- \Rightarrow \text{Class I: low silicon;}
\]
\[
\text{Patterns: } \text{Inputs Set } X^0 \Rightarrow \text{Class II: proper silicon;}
\]
\[
\text{Patterns: } \text{Inputs Set } X^+ \Rightarrow \text{Class III: high silicon.}
\]

Therefore, the next effort is focused on constructing a partly transparent 3-class classifier that could distinguish inputs sets \( X^- \), \( X^0 \) and \( X^+ \) as accurate as possible, and then provides suggestions for the blast furnace control based on inputs set \( X^0 \).
### Table 7: Relevant input variables of blast furnace (b)

| Variable name [Unit] | Symbol | Input variable |
|----------------------|--------|----------------|
| Blast temperature [°C] | \( x^{(3)} \) | \( q_0, q_{-1}, q_{-2}, q_{-3}, q_{-4} \) |
| Blast volume \([m^3/min]\) | \( x^{(4)} \) | \( q_0, q_{-1}, q_{-2}, q_{-3}, q_{-4} \) |
| Feed speed \([mm/h]\) | \( x^{(8)} \) | \( q_0, q_{-1}, q_{-2}, q_{-3}, q_{-4} \) |
| Gas permeability \([m^3/min-kPa]\) | \( x^{(11)} \) | \( q_0, q_{-1}, q_{-2}, q_{-3}, q_{-4} \) |
| Pulverized coal injection \([ton]\) | \( x^{(14)} \) | \( q_0, q_{-1}, q_{-2}, q_{-3}, q_{-4} \) |
| Sulfur content [wt%] | \( x^{(15)} \) | \( q_{-1} \) |
| Silicon content [wt%] | \( z \) | \( q_{-1} \) |

The variables with wave underlines are practical inputs fed into models.

### 5.2 Data collection and prior information acquisition

Experimental data is collected from two typical Chinese blast furnaces: One is a medium-sized blast furnace with the inner volume of about 2500 m³, and the other is a pint-sized blast furnace with the volume of about 750 m³, labeled blast furnace (a) and (b), respectively. The variables that are closely related to the hot metal silicon content are selected as the candidate inputs for modeling. Tables 6 and 7 present the selected 16 and 7 blast furnace variables, respectively. Note that the variables information from these two blast furnace are extended to include some lagged terms since there is usually 2 – 8h time delay for the blast furnace outputs responding to the inputs (Nurkkala et al., 2011). The sampling interval is about 1.5h for the blast furnace (a) while 2h for the blast furnace (b), so there are in all 86 and 27 relevant variables for them, respectively. However, we only use 42 feature variables for the blast furnace (a) and 9 for (b) as inputs to induce outputs (Gao et al., 2014), which have been marked out in Tables 6 and 7 with wave underlines. For these two blast furnace data sets, every set consists of 800 recordings, among which 700 points are set as training set and the remaining 100 recordings as testing set. As an example, Fig. 2 displays the silicon evolution for these two blast furnaces. According to the reported silicon infimum and supremum (Gao et al., 2014), \((0.4132; 0.8251)\) for the blast furnace (a) and \((0.3736; 0.8059)\) for the blast furnace (b), every silicon recording is labeled by \(-1\) (lower than the infimum), 0 (between the infimum and the supremum) or \(+1\) (higher than the supremum) to represent “low silicon”, “proper silicon” or “high silicon”. There are 104/213 “low silicon” samplings, 594/560 “proper silicon” samplings, and 96/23 “high silicon” samplings for the blast furnace (a)/(b). The data imbalance is quite obvious with “proper silicon” samplings in high proportion, more than 70%, which can be seen from the fact that most of samplings are distributed between two red dotted lines in Fig. 2.

The 3-class silicon classification tasks are performed by the one-against-all method (Vapnik et al., 1998) in which two binary classifiers need to be designed. For the blast furnace (a)/(b), the first one serves for classifying the “low/high silicon” samplings from other two classes while the second one is used to distinguish “proper silicon” from “high/low silicon” samplings. With the same experimental procedures as made for the benchmark examples, we firstly select 700 recordings randomly as the training set, and then apply the mining algorithm to acquiring linear priors for every blast furnace. The
Table 8: Linear Priors of Blast Furnaces (a) and (b)

| BF | Linear Priors                                                                 |
|----|------------------------------------------------------------------------------|
| (a) | $L_a^- : 0.6967q^{-1}z + 0.7174q^{-1}x^{(15)} - 0.2097 \leq 0 \Rightarrow x \in \mathcal{X}^-$ |
|    | $L_a^0 : 0.2837q^{-1}z - 0.9589q^0x^{(8)} + 0.6732 \leq 0 \Rightarrow x \in \mathcal{X}^0$ |
|    | $L_a^+ : 0.1865q^{-3}x^{(4)} - 0.9825q^{-3}x^{(12)} + 0.7627 \leq 0 \Rightarrow x \in \mathcal{X}^+$ |
| (b) | $L_b^- : 0.9553q^{-1}z + 0.2955q^{-3}x^{(14)} - 0.2307 \leq 0 \Rightarrow x \in \mathcal{X}^-$ |
|    | $L_b^0 : 0.5403q^{-1}x^{(8)} + 0.8415q^{-1}x^{(14)} - 0.6429 \leq 0 \Rightarrow x \in \mathcal{X}^0$ |
|    | $L_b^+ : -0.9983q^{-1}z - 0.0584q^{-2}x^{(11)} + 0.6459 \leq 0 \Rightarrow x \in \mathcal{X}^+$ |

results are presented in Table 8. To observe the boundaries more intuitively, they are displayed in Figs. 3 and 4 for the blast furnaces (a) and (b), respectively. Clearly, in the 1) and 3) sub-figures of these two figures, the “low silicon” and “high silicon” boundaries only separate several corresponding samplings from others. Especially for the blast furnace (b), there are only 3 “high silicon” samplings classified (see Fig. 4-3)). These results reveal that it is not easy to find good priors for those samplings with low proportion in the data set. Even if these priors can be captured, they are not necessarily to be able to play constructive role in improving model precision. The reason is that to guarantee some precision on these samplings with low proportion, the precision on those sampling with high proportion may reduce.

5.3 Experimental results and discussion

Incorporating the “low/high silicon” boundary into the first classifier and the “proper silicon” as well as “high/low silicon” boundaries into the second classifier, then we could get the degenerated $pTsm$-SVM like Eq. (30) for the blast furnace (a)/(b). In this model, we have some definite information on guiding how to classify samplings, as
suggested by the mined linear priors in Table 8. These pieces of information restrict the model structure and further affect the development of the solving algorithm, which leads to the degenerated \( pTsm-SVM \) is more transparent than the soft-margin SVM. Similar to what have been done on the benchmark examples, we have made 10 times random experiments on the blast furnace data with 700 samplings randomly generated as the training set for learning parameters and the remaining 100 samplings as the testing set for evaluating performance every time. Table 9 reports the detailed testing results for these two blast furnaces. For comparisons, we also provide the corresponding results produced by the black-box soft-margin SVM.

As can be seen from Tables 9, the incorporation of priors into the soft-margin SVM
models for the blast furnaces considered could improve the accuracy greatly, the ATA increasing from 68.5% to 73.4% for the blast furnace (a), and from 65.4% to 70.2% for the blast furnace (b). Moreover, the smaller standard deviations emerge, 4.3% v.s. 6.6% and 5.4% v.s. 8.4% for the blast furnaces (a) and (b), respectively. On the surface, these two indices both imply that the degenerated $pTsm$-SVM has better performance than the black-box soft-margin SVM. However, for the same reason as given in modeling the benchmark data sets, it is still uncertain to say that the former must have higher precision than the latter. We also make a paired Student’s t-test on the classification results yielded by these two models, and get the p-values to be 0.01 and 0.03 with the significant level of 0.05 for the blast furnace (a) and (b), respectively, shown in Table 9 too. These p-values results suggest that it is of high possibility for the degenerated $pTsm$-SVM to be more effective than the soft-margin SVM when they are applied to modeling the blast furnaces (a) and (b).

A further look at Table 9 might suggest that the precision improvement mainly results from the increase of the “proper silicon” samples to be classified right when linear priors are incorporated. For the blast furnace (a), it increases from 82.0% to 92.2%, and for the blast furnace (b) the result changes from 79.7% to 92.1%. This implies the outputs of the degenerated $pTsm$-SVM for these two blast furnace are trustworthy if they predict the label of “proper silicon”. Since the blast furnace operation is smooth during most of the time, which means the silicon content in hot metal to be often proper, the degenerated $pTsm$-SVM can work smoothly. However, there are some precision losses on classifying “low silicon” or “high silicon” samplings when incorporations take place, as expected from looking at Figs. 3 and 4. Especially for the blast furnace (a), the precision loss is from 55.0% to 37.6% in classifying “high silicon” samplings. In addition, we can notice from Table 9 that the distributions of the accuracy on “low silicon” samplings of the blast furnace (a) and “high silicon” samplings of the blast furnace (b) are obviously not symmetrical, which look very sparse. It indicates that the priors incorporation could enlarge the effect of samplings with high proportion on model identification, but weaken the influence of those samplings in low proportion. A great challenge will be encountered in classifying right the samplings in low proportion. The extreme imbalance of the blast furnace data make it difficult to conclude that the degenerated $pTsm$-SVM is more effective than the soft-margin SVM only based on classification accuracy. The unsymmetrical distribution of the classification accuracy also suggests it unfair to evaluate the stability of different algorithms by standard deviations. For these reasons, we employ other measures to compare these two models and try to make a fair evaluation for their applications to blast furnace. Cohen’s Kappa coefficient (Cano et al., 2013), denoted by $\kappa$ here, is an alternative statistic to evaluate the agreement between two raters. Since $\kappa$ considers the possibility of the agreement occurring by chance, it is more robust than the simple percent agreement calculation. The theoretical range of $\kappa$ is $[-1, 1]$ with −1 representing total disagreement while 1 representing total agreement. The larger $\kappa$ is, the smaller possibility of the agreement occurring by chance is. For the classification results in every blast furnace, we could calculate the Kappa statistic, the average of which over 10 times calculations is exhibited in Table 9. As can be seen, the average $\kappa$ will increase from 0.300 to 0.312 after the mined linear priors are incorporated into the black-box soft-margin SVM for the blast furnace (a) while from 0.151 to 0.164 for the blast furnace (b). The aver-
| Blast furnace | Experiments number | With priors, Accuracy (%) | Without priors, Accuracy (%) |
|---------------|--------------------|---------------------------|------------------------------|
|               |                    | total | low | proper | high | total | low | proper | high |
| (a)           | 1                  | 78.0  | 0.0 | 100.0  | 21.4 | 79.0  | 0.0 | 97.3   | 42.9 |
|               | 2                  | 74.0  | 0.0 | 95.9   | 36.4 | 63.0  | 0.0 | 75.3   | 72.7 |
|               | 3                  | 66.0  | 0.0 | 71.1   | 92.3 | 57.0  | 0.0 | 57.9   | 100.0|
|               | 4                  | 76.0  | 60.0| 87.3   | 35.7 | 73.0  | 60.0| 83.1   | 35.7 |
|               | 5                  | 74.0  | 0.0 | 98.7   | 8.3  | 74.0  | 0.0 | 100.0  | 8.3  |
|               | 6                  | 78.0  | 0.0 | 98.7   | 40.0 | 72.0  | 0.0 | 86.3   | 60.0 |
|               | 7                  | 69.0  | 0.0 | 97.1   | 11.1 | 70.0  | 0.0 | 95.7   | 22.2 |
|               | 8                  | 77.0  | 0.0 | 94.5   | 53.3 | 63.0  | 0.0 | 67.1   | 93.3 |
|               | 9                  | 74.0  | 0.0 | 97.1   | 35.3 | 70.0  | 0.0 | 84.3   | 64.7 |
|               | 10                 | 68.0  | 0.0 | 80.0   | 41.7 | 64.0  | 0.0 | 72.5   | 50.0 |
|               | Average            | 73.4  | 6.0 | 92.2   | 37.6 | 68.5  | 6.0 | 82.0   | 55.0 |
| ATA (total,%) |                    | 73.4 ±4.3 |     | 68.5 ±6.6 |     |
| p-value       |                    | 0.01  |     |        |      |
| average κ     |                    | 0.312 |     | 0.300  |      |
| (b)           | 1                  | 69.0  | 10.3| 98.5   | 0.0  | 68.0  | 6.9 | 98.5   | 0.0  |
|               | 2                  | 65.0  | 31.8| 79.5   | 0.0  | 57.0  | 90.9| 50.7   | 0.0  |
|               | 3                  | 82.0  | 33.3| 93.8   | 50.0 | 77.0  | 33.3| 87.5   | 50.0 |
|               | 4                  | 65.0  | 37.0| 80.9   | 0.0  | 62.0  | 33.3| 77.9   | 0.0  |
|               | 5                  | 73.0  | 4.0 | 98.6   | 0.0  | 77.0  | 36.0| 93.2   | 0.0  |
|               | 6                  | 72.0  | 14.8| 98.5   | 25.0 | 68.0  | 0.0 | 98.5   | 25.0 |
|               | 7                  | 71.0  | 4.2 | 98.6   | 0.0  | 70.0  | 0.0 | 98.6   | 0.0  |
|               | 8                  | 74.0  | 32.1| 94.2   | 0.0  | 52.0  | 78.6| 43.5   | 0.0  |
|               | 9                  | 66.0  | 19.4| 93.8   | 0.0  | 66.0  | 32.3| 87.5   | 0.0  |
|               | 10                 | 65.0  | 37.9| 84.4   | 0.0  | 57.0  | 62.1| 60.9   | 0.0  |
|               | Average            | 70.2  | 22.5| 92.1   | 3.75 | 65.4  | 37.3| 79.7   | 3.75 |
| ATA (total,%) |                    | 70.2 ±5.4 |     | 65.4 ±8.4 |     |
| p-value       |                    | 0.03  |     |        |      |
| average κ     |                    | 0.164 |     | 0.151  |      |
age $\kappa$ becomes larger when the transparency of the black-box model is enhanced for both blast furnaces. Hence, the accuracy increase from the soft-margin SVM to the degenerated $pTsm$-SVM most likely results from the incorporation of the mined blast furnace linear priors but not from chance. This together with the $p$-values of $t$-test calculated with respect to these two models makes it sure that the degenerated $pTsm$-SVM is more effective than the black-box soft-margin SVM. Additionally, in view of the unsymmetrical distribution of the classification accuracy, the box-plot is employed as an additional measure to analyze the stability of two models. Fig. 5 displays the box-plots with respect to the accuracy of two models at 10 times random experiments. Clearly, when the transparency of the black-box soft-margin is added, the prediction accuracy is more concentrated for both blast furnaces. Moreover, the whole accuracy range of the degenerated $pTsm$-SVM looks higher than that of the black-box soft-margin SVM. As a result, the former model exhibits more stable performance than the latter one, which supports the conclusion drawn from the standard deviations.

The above analysis indicates that when applied to the studied blast furnaces, the degenerated $pTsm$-SVM can outperform the soft-margin SVM in both of transparency and accuracy. This result is of large practical significance. Due to great quantity of production in blast furnace process, a slight improvement in model precision may lead to considerable profit.

6 Conclusions and points of possible future research

This paper has presented a theoretical and experimental study on the incorporation of data-based prior information into black-box soft-margin SVM model for transparency
and precision enhancement. The main contribution includes: i) propose a concise and practical algorithm to mine linear prior information from data set; ii) for the soft-margin SVM with priors incorporated, develop an equivalent model that seems to have the same structure as “pure” black-box SVM so that the existent commercial software packages can be directly utilized to find solutions; iii) provide a feasible solution to enhance both of the transparency and precision of blast furnace black-box models.

Despite the good performance exhibited by the degenerated $p\text{Tsm}$-SVM, there are still great rooms for the model to be improved. The most urgent task maybe achieves the theoretical support that the degenerated $p\text{Tsm}$-SVM is able to converge the true one, i.e., the corresponding Bayes model. The estimation of the convergence rate is also an important investigation point. In addition, more concerns may be thrown towards reducing the training time of the transparent model. The current framework is quite time-consuming since the incorporation of priors would introduce more parameters needed to be studied. The next effort is to estimate the theoretical optimal values of these parameters so that the transparent model has the same number of parameters to be trained with the corresponding black-box model.

**Acknowledgments**

The financial supports from the National Natural Science Foundation of China under Grant No. 11671418, 11271326 and 61611130124, and the Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20130101110040 are highly appreciated.

**References**

Abu-Mostafa, Y. S. (1990). Learning from hints in neural networks. *Journal of Complexity, 6*(2), 192 – 198.

Aguirre, L. A., Lopes, R. A., Amaral, G. F., & Letellier, C. (2004). Constraining the topology of neural networks to ensure dynamics with symmetry properties. *Physical Review E, 69*(2), 026701.

Barakat, N., & Bradley, A. P. (2010). Rule extraction from support vector machines: a review. *Neurocomputing, 74*(1), 178 – 190.

Bhattacharya, T. (2005). Prediction of silicon content in blast furnace hot metal using partial least squares (PLS). *ISIJ International, 45*(12), 1943 – 1945.

Borges, R. V., Garcez, A. D. A., & Lamb, L. C. (2011). Learning and representing temporal knowledge in recurrent networks. *IEEE Transactions on Neural Networks, 22*(12), 2409 – 2421.

Cano, A., Zafra, A., & Ventura, S. (2013). Weighted data gravitation classification for standard and imbalanced data. *IEEE Transactions on Cybernetics, 43*(6), 1672 – 1687.
Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27.

Chen, P. H., Lin, C. J., & Schölkopf, B. (2005). A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry, 21*(2), 111–136.

Chen, S., Wolfgang, A., Harris, C. J., & Hanzo, L. (2008). Symmetric RBF classifier for nonlinear detection in multiple-antenna-aided systems. *IEEE Transactions on Neural Networks, 19*(5), 737–745.

Chorowski, J., & Zurada, J. M. (2011). Extracting rules from neural networks as decision diagrams. *IEEE Transactions on Neural Networks, 22*(12), 2435–2446.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*, Cambridge University Press.

Daniels, H., & Velikova, M. (2010). Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks, 21*(6), 906–917.

Dugas C., Bengio Y., Blisle F., Nadeau C., & Garcia R. (2009) Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research, 10*(1), 1239–1262.

Frasconi, P., Gori, M., Maggini, M., & Soda, G. (1991, July). An unified approach for integrating explicit knowledge and learning by example in recurrent networks. *Proceedings of the International Joint Conference on Neural Networks* (Vol. 1, pp. 811-816). IEEE.

Gao, C., Chen, J., Zeng, J., Liu, X., & Sun, Y. (2009). A chaos-based iterated multistep predictor for blast furnace ironmaking process. *AIChE Journal, 55*(4), 947–962.

Gao, C., Ge, Q., & Jian, L. (2014). Rule extraction from fuzzy-based blast furnace SVM multiclassifier for decision-making. *IEEE Transactions on Fuzzy Systems, 22*(3), 586–596.

Gao, C., Jian, L., & Luo, S. (2012). Modeling of the thermal state change of blast furnace hearth with support vector machines. *IEEE Transactions on Industrial Electronics, 59*(2), 1134–1145.

Grumiller, D., McNees, R., & Zonetti, S. (2012). Black holes in the conical ensemble. *Physical Review D, 86*(12), 124043.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hu, B. G., Wang, Y., Yang, S. H., & Qu, H. B. (2007). How to add transparency to artificial neural networks. *Pattern Recognition and Artificial Intelligence, 20*(1), 72–84.

Hu, M., Chen, Y., & Kwok, J. T. Y. (2009). Building sparse multiple-kernel SVM classifiers. *IEEE Transactions on Neural Networks, 20*(5), 827–839.
Hu, Z., Yang, Z., Salakhutdinov, R., & Xing, E. P. (2016). Deep neural networks with massive learned knowledge. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1670-1679).

Huynh, T. Q., & Reggia, J. A. (2011). Guiding hidden layer representations for improved rule extraction from neural networks. *IEEE Transactions on Neural Networks, 22*(2), 264 – 275.

Jian, L., Gao, C., & Xia, Z. (2012). Constructing multiple kernel learning framework for blast furnace automation. *IEEE Transactions on Automation Science and Engineering, 9*(4), 763 – 777.

Jindal, A., Pujari, S., Sandilya, P., & Ganguly, S. (2007). A reduced order thermo-chemical model for blast furnace for real time simulation. *Computers & Chemical Engineering, 31*(11), 1484 – 1495.

Khosrowabadi, R., Quek, C., Ang, K. K., & Wahab, A. (2014). ERNN: A biologically inspired feedforward neural network to discriminate emotion from EEG signal. *IEEE Transactions on Neural Networks and Learning Systems, 25*(3), 609 – 620.

Kuhn, H. W., & Tucker, A. W. Nonlinear programming, 1951. *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics* (pp. 481-492).

Lauer, F., & Bloch, G. (2008a). Incorporating prior knowledge in support vector regression. *Machine Learning, 70*(1), 89 – 118.

Lauer, F., & Bloch, G. (2008b). Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing, 71*(7), 1578 – 1594.

Liu, X., Wang, L., Yin, J., Zhu, E., & Zhang, J. (2013). An efficient approach to integrating radius information into multiple kernel learning. *IEEE Transactions on Cybernetics, 43*(2), 557 – 569.

Martens, D., Baesens, B. B., & Van Gestel, T. (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 178 – 191.

Mangasarian, O. L., Shavlik, J. W., & Wild, E. W. (2004). Knowledge-based kernel approximation. *Journal of Machine Learning Research, 5*(Sep), 1127 – 1141.

Mangasarian, O. L., & Wild, E. W. (2007). Nonlinear knowledge in kernel approximation. *IEEE Transactions on Neural Networks, 18*(1), 300 – 306.

Mangasarian, O. L., & Wild, E. W. (2008). Nonlinear knowledge-based classification. *IEEE Transactions on Neural Networks, 19*(10), 1826 – 1832.

Minin A., Velikova M., Lang B., & Daniels H. (2010). Comparison of universal approximators incorporating partial monotonicity by structure. *Neural Networks, 23*(4), 471 – 475.
Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 86(11), 2196 – 2209.

Nogami, H., Chu, M., & Yagi, J. I. (2005). Multi-dimensional transient mathematical simulator of blast furnace process based on multi-fluid and kinetic theories. *Computers & Chemical Engineering*, 29(11), 2438 – 2448.

Nurkkala, A., Pettersson, F., & Saxén, H. (2011). Nonlinear modeling method applied to prediction of hot metal silicon in the ironmaking blast furnace. *Industrial & Engineering Chemistry Research*, 50(15), 9236 – 9248.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), 1481 – 1497.

Qu, Y. J., & Hu, B. G. (2011). Generalized constraint neural network regression model subject to linear priors. *IEEE Transactions on Neural Networks*, 22(12), 2447 – 2459.

Saxén, H., Gao, C., & Gao, Z. (2013). Data-driven time discrete models for dynamic prediction of the hot metal silicon content in the blast furnace: A review. *IEEE Transactions on Industrial Informatics*, 9(4), 2213 – 2225.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* Cambridge, MA: MIT Press.

Towell, G. G., & Shavlik, J. W. (1994). Knowledge-based artificial neural networks. *Artificial Intelligence*, 70(1-2), 119 – 165.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67 – 82.

Wu, S. H., Lin, K. P., Chien H. H., Chen, C. M., & Chen, M. S. (2013). On generalizable low false-positive learning using asymmetric support vector machines. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 1083 – 1096.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

Velikova M., H. Daniels H., & Feelders A. (2006). Mixtures of monotone networks for prediction. *International Journal of Computational Intelligence*, 3(3), 204 – 214.

Xu, X., Tsang, I. W., & Xu, D. (2013). Soft margin multiple kernel learning. *IEEE transactions on Neural Networks and Learning Systems*, 24(5), 749 – 761.

Yuan, Y., Lian, X., Chen, L., Sun Y., & Wang, G. (2016). RSkNN: kNN search on road networks by incorporating social influence. *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1575 – 1588.

Zhao, X., Li, X., & Zhang, Z. (2015). Joint structural learning to rank with deep linear feature learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(10), 2756 – 2769.
