Enhancing Interpretability of Black-box Soft-margin SVM by Integrating Data-based Priors

Shaohan Chen, Chuanhou Gao, *Senior Member, IEEE,* and Ping Zhang

Abstract—The lack of interpretability often makes 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 interpretability. The concept and incorporation mechanism of data-based prior information are successively developed, based on which the interpretable or partly interpretable 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 interpretability enhancement strategy is applied to eight benchmark examples for effectiveness exhibition.

Index Terms—Soft-margin SVM, black-box, interpretability, data-based, prior information

I. INTRODUCTION

DEVELOPMENT of black-box modeling techniques, like support vector machine (SVM), neural networks, etc., has shown rather rapid in the past decades. 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 [1], black hole [2], integrated industrial processes [3] 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 on the operating principle of black-box models. The loss of interpretability makes it impossible to explain the model outputs as comprehensive knowledge, and neither to improve the model
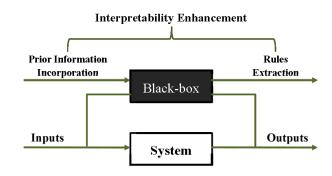
Fig. 1. Interpretability enhancement ways for the black-box model.

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 desired and necessary. It thus becomes quite significant to investigate ways of enhancing interpretability of black-box models, which has been the goals of DARPA's currently ongoing Explainable AI project [4]. Of particular interest is the algorithm called to Local Interpretable Model-Agnostic Explanations (LIME). LIME seeks to extract expressivity from black-box models.

There are two ways to make a black-box model more interpretable. One way is from the input port of the black box to integrate prior information into the model [5]–[7], the other way is from the output port to extract comprehensive rules from the model [8]–[10]. A schematic diagram to display these two interpretability enhancement 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. [11]. 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 [12] that states all algorithms perform the same when averaged over different problems. Lauer et al. [13] 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 [14] in the case that the data size is finite. More importantly, the inclusion of priors into the model can add both of interpretability and precision. Furthermore, unlike rule extraction only giving interpretability

in knowledge reasoning, the prior information may refer to any aspect about the problem, like the structure, parameters, etc., so it can provide interpretability from every aspect [15].

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 [16], symmetry [17], monotonicity [18], [19], boundary constraint/input domain [5], [20]–[22], etc., and knowledge on the data [13], 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 [23], [24]; 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 [5] into the model; the boundary constraint/input domain is embedded as equality or inequality constraint [20]–[22]. For the information of non-persistent exciting data, Niyogi et. al. [14] 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 [25]. 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 classification method may be based on the black-box models that need to be enhanced interpretability. 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. [26] 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 [27] and monotone neural network [28], [29], where partial monotonicity or monotonicity is incorporated by structure. Dugas et al. [30] 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. [15] 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. [31] 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 [23]–[25] or in the form of inequality constraint [5], [20]–[22]. More details may be found in the review paper [13].

This work continues to enrich the methods of integrating the prior information into the black-box model. We take the soft-margin SVM [32], [33] as an example of black-box models to be incorporated with the priors for interpretability 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 interpretable soft-margin SVM model. In the model construction, we fully consider the correctness of the mined linear priors, and design the interpretable model as the balance of maximizing the margin and minimizing the errors of the priors. In the model solving process, we rewrite the interpretable model to have the same structure as the pure black-box soft margin SVM so that the common SVM software package [34] can be used directly. The mined linear priors are embedding into the black-box model in the form of inequality constraints [20]–[22], which add interpretability 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 to evaluate the performance of the proposed interpretability enhancement method.

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 interpretability enhancement 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 $\mathbf{z}_a = (z_{a_1}, \cdots, z_{a_n})^\top$ and $\mathbf{z}_b = (z_{b_1}, \cdots, z_{b_n})^\top \in \mathbb{R}^n$ the inequality $\mathbf{z}_a \geq \mathbf{z}_b$ means $z_{a_i} \geq z_{b_i}$, $\forall i$. More notations may be found in Table 1.

II. SOFT-MARGIN SUPPORT VECTOR MACHINE

Consider a binary classification problem. Let $\mathbf{x} \in \mathbb{R}^n$, $y \in \{+1, -1\}$ represent the input pattern and output pattern,

TABLE I
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 $\boldsymbol{\alpha}$
G_1, G_2	Two functions representing the terms related to the kernel function and the Lagrangian multipliers, respectively
$\mathbf{k}(\cdot, \cdot)$	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
$\bar{\mathbb{R}}^n_+$	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 Γ^+ and Γ^- 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 Γ^+ and Γ^-, respectively
$\mathbf{x}^-, \mathbf{x}^-$	Sampling input to render "positive/negative class" prior information
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$	Lagrangian multiplier vectors
$\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}$	Improved version of Lagrangian multiplier vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively
Γ^+, Γ^-	The subset of $\mathbb{R}^n$ containing sampling input $\mathbf{x}^-, \mathbf{x}^-$, respectively
$\mathcal{L}(\boldsymbol{\alpha})$	The dual target margin
$\mathscr{L}^+_{ij}, \mathscr{L}^-_{ij}$	Labeling linear "positive and negative class" prior information in the plane $X^{(i)}OX^{(j)}$ composed of two features $x^{(i)}$ and $x^{(j)}$
$\boldsymbol{\zeta}, \boldsymbol{\varsigma}$	Slack variable vectors to measure incorrectness of the mined "positive and negative class" prior information, respectively
θ	Kernel slack variable
$\lambda_1, \lambda_2, \lambda_3$	Counterbalance constants
$\boldsymbol{v}_+, \boldsymbol{v}_-$	r^+ and r^- dimensional nonnegative real vectors, respectively
Φ	Feature map from $\mathbb{R}^n$ to high dimensional feature space
$\mathbb{0}_n, \mathbb{1}_n$	n-dimensional vector with all entries equal to 0 and 1, respectively
s.t.	The abbreviation of "subject to"

respectively. The ν-SVM [32] aims at solving the optimization problem

$$\min_{\mathbf{w}, b, \rho \geq 0, \xi_i \geq 0} \quad -\nu\rho + \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{1}{N}\sum_{i=1}^{N}\xi_i,$$
$$\text{s.t.} \quad y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i) + b) \geq \rho - \xi_i, \forall i, \tag{1}$$

where $\mathbf{w}$ is the normal vector in an imaginary high-dimensional feature space, $N \in \mathbb{R}$ is the number of samplings, $b \in \mathbb{R}$ is the offset, $\Phi(\cdot)$ is a high-dimensional feature project from $\mathbb{R}^n$ to the feature space, and $\nu \in [0,1]$ represents the upper bound on the fraction of training errors or the lower bound on the fraction of support vectors. If $\mathbf{w}^\top\Phi(\mathbf{x}_i) + b \geq \rho - \xi_i$ for a $\mathbf{x}_i$ then this $\mathbf{x}_i$ belongs to the class of $y_i = +1$ while if $\mathbf{w}^\top\Phi(\mathbf{x}_i) + b \leq -\rho + \xi_i$ then $\mathbf{x}_i$ belongs to the class of $y_i = -1$. Here, ξ_i $(i = 1, \cdots, N)$ are non-negative slack variables to measure the degree of misclassification of the data $\mathbf{x}_i$. By utilizing Lagrangian multipliers $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)^\top \in \bar{\mathbb{R}}^N_+$ and further making dual transformation, the optimization problem of Eq. (1) can be written as

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}(\boldsymbol{\alpha}),$$
$$\text{s.t.} \quad \boldsymbol{\alpha}^\top\mathbf{y} = 0, \mathbb{0}_N \leq \boldsymbol{\alpha} \leq \frac{1}{N}\mathbb{1}_N, \boldsymbol{\alpha}^\top\mathbb{1}_N \geq \nu, \tag{2}$$

where $\mathcal{L}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$, and $k(\mathbf{x}_i, \mathbf{x}_j)$ is the well-known kernel function [25] defined by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^\top\Phi(\mathbf{x}_j). \tag{3}$$

The decision function for ν-SVM takes the form of

$$f(\mathbf{x}) = \text{sign}\Big(\sum_{\alpha_i > 0}^{i=1 \sim N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\Big). \tag{4}$$

The optimization problem of Eq. (2) may be further relaxed by introducing the nonnegative kernel slack variable θ [33], defined by the difference of the target margin τ and the above dual target $\mathcal{L}(\boldsymbol{\alpha})$, i.e.,

$$\theta = \tau - \mathcal{L}(\boldsymbol{\alpha}). \tag{5}$$

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

$$\min_{\tau, \boldsymbol{\alpha} \in \mathcal{A}, \theta \geq 0} \quad -\tau + \lambda_1 l_1(\theta),$$
$$\text{s.t.} \quad \mathcal{L}(\boldsymbol{\alpha}) \geq \tau - \theta, \tag{6}$$

where $\mathcal{A} = \{\boldsymbol{\alpha} | \boldsymbol{\alpha}^\top\mathbf{y} = 0, \mathbb{0}_N \leq \boldsymbol{\alpha} \leq \frac{1}{N}\mathbb{1}_N, \boldsymbol{\alpha}^\top\mathbb{1}_N \geq \nu\}$, the parameter $\lambda_1 \in \mathbb{R}_+$ acts as a counterbalance, and $l_1(\cdot)$ represents any loss function. Note that the introduction of θ means that the margin of classification changes from the hard one τ to the soft one $\tau - \theta$. We thus refer to Eq. (6) as soft-margin SVM in the following. However, the solutions will keep unchanged for these two optimization problems.

Proposition 1: Soft-margin SVM of Eq. (6) has the same solutions as ν-SVM of Eq. (2).

Proof: This result is a special case of *Proposition 2* in the paper [33]. See the detailed proof there. ∎

III. DATA-BASED PRIOR INFORMATION

A. Prior Information

Prior information incorporated into black-box models will add a high degree of interpretability. 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 [14]. Here, prior information is defined as follows.

Definition 1: [6]. 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 techniques. These logical implications are quite like the nonlinear prior knowledge proposed by Mangasarian et. al. [20]–[22]. In the following, we will shortly introduce the formulation of that kind of prior knowledge [22] within the framework of SVM black-box model.

Let the training examples be $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, Γ^+, $\Gamma^- \subseteq \mathbb{R}^n$ and $g^+ : \Gamma^+ \to \mathbb{R}^{r^+}$, $g^- : \Gamma^- \to \mathbb{R}^{r^-}$, where r^+ and r^- are two positive integers, then the nonlinear prior information is expressed as

$$g^+(\mathbf{x}^+) \leq \mathbb{0}_{r+} \Longrightarrow \sum_{i=1}^N \alpha_i y_i k(\mathbf{x}^+, \mathbf{x}_i) + b \geq b^*, \quad \forall \mathbf{x}^+ \in \Gamma^+ \tag{7}$$

to identify positive class points $y = +1$, and

$$g^-(\mathbf{x}^-) \leq \mathbb{0}_{r-} \Longrightarrow \sum_{i=1}^N \alpha_i y_i k(\mathbf{x}^-, \mathbf{x}_i) + b \leq -b^*, \forall \mathbf{x}^- \in \Gamma^- \tag{8}$$

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. (7), means that

$$\left. \begin{array}{l} g^+(\mathbf{x}^+) \leq \mathbb{0}_{r+}, \\ \mathbf{z}^\top \mathbf{k}(\mathbf{x}^+) + b < b^* \end{array} \right\} \quad \text{has no solution,} \tag{9}$$

where $\mathbf{k}(\mathbf{x}^+) = (k(\mathbf{x}^+, \mathbf{x}_1), \cdots, k(\mathbf{x}^+, \mathbf{x}_N))^\top$ and $\mathbf{z} = (\alpha_1 y_1, \cdots, \alpha_N y_N)^\top$. Moreover, if Γ^+ is a convex subset of $\mathbb{R}^n$, and $g^+(\mathbf{x}^+)$ and $\mathbf{k}(\mathbf{x}^+)$ are convex on Γ^+, the following result holds.

Lemma 1: [22]. The prior information expressed as Eq. (7) or Eq. (9) is equivalent to the result that there exists $\boldsymbol{v}_+ \in \bar{\mathbb{R}}_+^{r^+}$ such that

$$\mathbf{z}^\top \mathbf{k}(\mathbf{x}^+) + b - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}^+) \geq 0. \tag{10}$$

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

Lemma 2: [22]. The prior information expressed as Eq. (14) is equivalent to the result that there exists $\boldsymbol{v}_- \in \bar{\mathbb{R}}^{r^-}$ such that

$$-\mathbf{z}^\top \mathbf{k}(\mathbf{x}^-) - b - b^* + \boldsymbol{v}_-^\top g^-(\mathbf{x}^-) \geq 0. \tag{11}$$

The work of Mangasarian et. al. [20]–[22] successfully converted the prior information in the form of logical implication into the corresponding nonlinear inequality constraint. Obviously, the latter is more easily to be incorporated into black-box models. It has been shown in some numerical experiments that this incorporation can improve greatly precision of SVM models [20]–[22].

B. 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 acquirement of prior information is thus modeled by minimizing the loss of its incorrectness.

We also consider a binary classification problem for the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Let $h(\cdot)$ represent a classifier rendered by any black-box model, in the case of SVM model which takes the form of $h(\cdot) = f(\cdot) = \sum_{i=1}^N \alpha_i y_i k(\cdot, \mathbf{x}_i) + b$. The classification occurs according to

$$y = \left\{ \begin{array}{ll} +1, & h(\mathbf{x}) \geq b^*, \\ -1, & h(\mathbf{x}) \leq -b^*. \end{array} \right. \tag{12}$$

Let $g^+(\cdot) \leq \mathbb{0}_{r+}$ and $g^-(\cdot) \leq \mathbb{0}_{r-}$ be the positive and negative class functions that need to be mined, respectively, then the positive class and negative class prior information are expressed as

$$g^+(\mathbf{x}^+) \leq \mathbb{0}_{r+} \Longrightarrow h(\mathbf{x}^+) \geq b^*, \quad \forall \mathbf{x}^+ \in \Gamma^+, \tag{13}$$

and

$$g^-(\mathbf{x}^-) \leq \mathbb{0}_{r-} \Longrightarrow h(\mathbf{x}^-) \leq -b^*, \quad \forall \mathbf{x}^- \in \Gamma^-. \tag{14}$$

From the assumption A2, these two pieces of knowledge may be not true. We thus introduce nonnegative slack variables ζ and ς to measure their incorrectness, respectively. Based on **Lemmas** 1 and 2, they are defined by

$$\zeta = b^* - [h(\mathbf{x}^+) + \boldsymbol{v}_+^\top g^+(\mathbf{x}^+)], \quad \forall \mathbf{x}^+ \in \Gamma^+, \tag{15}$$

and

$$\varsigma = b^* - [\boldsymbol{v}_-^\top g^-(\mathbf{x}^-) - h(\mathbf{x}^-)], \quad \forall \mathbf{x}^- \in \Gamma^-. \tag{16}$$

The loss induced by these two error variables may be evaluated by any loss function, similar to penalizing θ in Eq. (6) for soft-margin SVM [33]. Denote the loss function by $l_2(\cdot)$ and $l_3(\cdot)$ for ζ and ς, respectively, then the penalty to the incorrectness may create the following optimization problem

$$\min_{\zeta, \varsigma \geq 0} \lambda_2 l_2(\zeta) + \lambda_3 l_3(\varsigma), \tag{17}$$

$$\text{s.t.} \quad h(\mathbf{x}^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}^+) + \zeta \geq 0, \quad \forall \mathbf{x}^+ \in \Gamma^+,$$

$$h(\mathbf{x}^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}^-) - \varsigma \leq 0, \quad \forall \mathbf{x}^- \in \Gamma^-,$$

where λ_2, $\lambda_3 \in \mathbb{R}_+$ are constants, called counterbalance, like λ_1 in Eq. (6). The solution may suggest two pieces of prior information: 1) If $g^+(\mathbf{x}^+) \leq \mathbb{0}_{r+}$, then $\mathbf{x}^+$ has the class label $y = +1$; 2) If $g^-(\mathbf{x}^-) \leq \mathbb{0}_{r-}$, then $\mathbf{x}^-$ has the class label $y = -1$. Thus, we have the following result.

Proposition 2: For any system with input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{+1, -1\}$, assume $h(\mathbf{x})$ defined by Eq. (12) to be a decision function for addressing binary classification problem of this system. Then the prior information: 1) $g^+(\mathbf{x}^+) \leq \mathbb{0}_{r+} \Longrightarrow y = +1$ and 2) $g^-(\mathbf{x}^-) \leq \mathbb{0}_{r-} \Longrightarrow y = -1$ can be corrected furthest by solving the optimization problem of Eq. (17).

Proof: The result is straightforward from **Lemmas** 1 and 2. ∎

The incorrectness of the prior information mined through data mining techniques is fully considered in the above

optimization problem, 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. (17), then the acquired prior information is called data-based prior information.

Remark 1: The modification of data-based prior information depends strongly on the black-box model for which it serves. In practical operation, this modification process and the black-box model training should be implemented synchronously. A naive idea for this purpose is to merge their object functions and constraints, respectively, to form one optimization problem, which simultaneously finishes the incorporation of prior information into the black-box model.

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 interpretability 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 interpretability 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 interpretability 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 interpretability and precision of black-box models.

IV. 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 interpretability and also maintaining high precision. The classifier in Eq. (17) thus takes the form of $h(\cdot) = f(\cdot)$.

For soft-margin SVM model of Eq. (6), it requests to minimize the margin error while for data-based prior information model of Eq. (17) minimizing the priors incorrectness. Their incorporation is naturally made through minimizing the sum of two objective functions, as said in *Remark 1*, i.e., making a trade-off between a large margin and small error penalties, which creates the following optimization problem

$$\min_{\tau, \boldsymbol{\alpha} \in \mathcal{A}, \theta \geq 0, \zeta \geq 0, \varsigma \geq 0} -\tau + \boldsymbol{\lambda}^\top \boldsymbol{l}, \qquad (18)$$
$$\text{s.t.} \quad \mathcal{L}(\boldsymbol{\alpha}) \geq \tau - \theta,$$
$$f(\mathbf{x}^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}^+) + \zeta \geq 0, \quad \forall \, \mathbf{x}^+ \in \Gamma^+,$$
$$f(\mathbf{x}^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}^-) - \varsigma \leq 0, \quad \forall \, \mathbf{x}^- \in \Gamma^-,$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^\top$ and $\boldsymbol{l} = (l_1(\theta), l_2(\zeta), l_3(\varsigma))^\top$.

Remark 2: Although the data-based priors, i.e., the last two constraints in Eq. (18), are mined from the data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, which means these constraints naturally hold for

some of the data points, they have to hold for all feasible (infinitely many) points in region of the feature space defined by the prior knowledge constraints. Namely, the prior knowledge constraints in Eq. (18) are semi-infinite [35] in the variables $\mathbf{x}^+$ and $\mathbf{x}^-$, and the optimization problem of Eq. (18) is a semi-infinite optimization problem.

For a semi-infinite optimization problem, if the SVM model and the prior knowledge constraints are linear [36], the problem is easy to solve since the semi-infinite variables may be removed by applying theorems of the alternative [37]. However, when any of them is nonlinear, this leads to a semi-infinite nonlinear optimization problem, which is very difficult to solve as it is NP-Hard. To circumvent this issue, Mangasarian and Wild [22] discretize the semi-infinite variables into a uniform mesh, which may convert a semi-infinite program [35] with an infinite number of constraints into a finite mathematical program. Note that the soft margin SVM is nonlinear and we also think the data-based priors to be nonlinear so that they are easy to be generalized. Therefore, Eq. (18) is a semi-infinite nonlinear optimization problem. We thus adopt the same discretizing technique as presented in [22] to convert it into a finite mathematical program, i.e.,

$$\min_{\tau, \boldsymbol{\alpha} \in \mathcal{A}, \theta \geq 0, \zeta \geq \mathbb{0}_{t+}, \varsigma \geq \mathbb{0}_{t-}} -\tau + \boldsymbol{\lambda}^\top \boldsymbol{l}, \qquad (19)$$
$$\text{s.t.} \quad \mathcal{L}(\boldsymbol{\alpha}) \geq \tau - \theta,$$
$$f(\mathbf{x}_j^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}_j^+) + \zeta_j \geq 0, j = 1, \cdots, q^+,$$
$$f(\mathbf{x}_h^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}_h^-) - \varsigma_h \leq 0, h = 1, \cdots, q^-,$$

where q^+ and q^- are the numbers of finite meshes of points for positive class prior and negative class prior, respectively, and

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

This model will act as the benchmark model that can enhance the interpretability but without loss of precision of the black-box soft-margin SVM. We refer to it and its equivalent models as partly interpretable 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. (19) is the same as that of the following optimization problem

$$\min_{\boldsymbol{\alpha}} \max_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \quad F_1(\boldsymbol{\alpha}) + F_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) + F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma}) + \boldsymbol{\lambda}^\top \boldsymbol{\ell}, \qquad (20)$$

where

$$\begin{cases} F_1(\boldsymbol{\alpha}) = -\mathcal{L}(\boldsymbol{\alpha}), \\ F_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{j=1}^{q^+} \beta_j [f(\mathbf{x}_j^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}_j^+)], \\ F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{h=1}^{q^-} \gamma_h [f(\mathbf{x}_h^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}_h^-)], \end{cases} \qquad (21)$$

and

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

with $l_1'(\cdot), l_2'(\cdot), l_3'(\cdot)$ representing the corresponding derivatives with respect to their respective argument. The decision variables $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{N^+})^\top \geq \mathbb{0}_{N^+}$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{N^-})^\top \geq \mathbb{0}_{N^-}$ are the Lagrangian multipliers.

Proof: The conversion from Eq. (19) to Eq. (20) is easily realized by constructing the Lagrangian function of the former, and then utilizing the corresponding KKT conditions. ∎

Remark 3: The optimization model of Eq. (20) will degenerate to be

$$\min_{\boldsymbol{\alpha}} \max_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \quad F_1(\boldsymbol{\alpha}) + F_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) + F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \tag{22}$$

if all the loss functions are selected as the hinge loss, i.e., $l(\cdot) = \max(0, \cdot)$ [33]. In this case, the decision variables satisfy $\mathbb{0}_{q^+} \leq \boldsymbol{\beta} \leq \lambda_2 \mathbb{1}_{q^+}$, $\mathbb{0}_{q^-} \leq \boldsymbol{\gamma} \leq \lambda_3 \mathbb{1}_{q^-}$ and $\boldsymbol{\alpha} \in \mathcal{A}$.

Remark 4: For the degenerated *pTsm*-SVM in Eq. (22), the objective consists of three functions, $F_1(\boldsymbol{\alpha})$, $F_2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, 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 5: Besides enhancing interpretability, 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 $F_1(\boldsymbol{\alpha})$, $F_2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ in Eq. (21) reveals that $F_1(\boldsymbol{\alpha})$ has different structure regarding $\boldsymbol{\alpha}$ from $F_2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $F_3(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. This may lead to large difficulty in utilizing the common SVM software package, like LibSVM [34]. To be applied conveniently, we reformulate the benchmark model of Eq. (19) for *pTsm*-SVM as follows.

Proposition 4: The optimization problem

$$\min_{\tau, \boldsymbol{\alpha} \in \mathcal{A}, \theta \geq 0, \boldsymbol{\zeta} \geq \mathbb{0}_{q^+}, \boldsymbol{\varsigma} \geq \mathbb{0}_{q^-}} -\tau + \boldsymbol{\lambda}^\top \boldsymbol{l}, \tag{23}$$

$$\text{s.t.} \quad \mathcal{L}(\boldsymbol{\alpha}) \geq \tau - \theta,$$
$$\alpha_j^+ y_j^+ [f(\mathbf{x}_j^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}_j^+) + \zeta_j] \geq 0, \forall j,$$
$$\alpha_h^- y_h^- [f(\mathbf{x}_h^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}_h^-) - \varsigma_h] \geq 0, \forall h$$

is equivalent to *pTsm*-SVM in Eq. (19), where α_j^+ and α_h^- are entries in $\boldsymbol{\alpha}$ 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. (19). Therefore, the result is true. ∎

Consider all losses in Eq.(23) induced by the hinge loss function, then we have the proposition as follows.

Proposition 5: The following optimization problem

$$\min_{\boldsymbol{\alpha} \in \mathcal{A}} \max_{\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}} \quad F_1(\boldsymbol{\alpha}) + \widetilde{F}_2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}) + \widetilde{F}_3(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\gamma}}) \tag{24}$$

shares the same solution with *pTsm*-SVM in Eq. (23) if all the losses induced by the slack variables θ, $\boldsymbol{\zeta}$ and $\boldsymbol{\varsigma}$ obey the rule of the hinge loss, where

$$\begin{cases} \widetilde{F}_2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}) = -\sum_{j=1}^{q^+} \widetilde{\beta}_j \alpha_j^+ [f(\mathbf{x}_j^+) - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}_j^+)], \\ \widetilde{F}_3(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\gamma}}) = \sum_{h=1}^{q^-} \widetilde{\gamma}_h \alpha_h^- [f(\mathbf{x}_h^-) + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}_h^-)], \end{cases}$$

$\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$ are the corresponding Lagrangian multiplier vectors.

Proof: Eq. (24) is the dual form of Eq. (23) under the given conditions, and they thus have the same solution. ∎

Remark 6: The decision variables $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$ in Eq. (24) have bounds as $\forall j$, $0 \leq \widetilde{\beta}_j \leq \frac{\lambda_2}{\alpha_j^+}$ and $\forall h$, $0 \leq \widetilde{\gamma}_h \leq \frac{\lambda_3}{\alpha_h^-}$.

For the convenience of using the common software package, further rewriting the optimization objective of Eq. (24) yields

$$\min_{\boldsymbol{\alpha} \in \mathcal{A}} \max_{\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}} \quad G_1(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}) - G_2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}), \tag{25}$$

where

$$\begin{cases} G_1(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}) = F_1(\boldsymbol{\alpha}) - \sum_{j=1}^{q^+} \widetilde{\beta}_j \alpha_j^+ \mathbf{z}^\top \mathbf{k}(\mathbf{x}_j^+) \\ \qquad\qquad + \sum_{h=1}^{q^-} \widetilde{\gamma}_h \alpha_h^- \mathbf{z}^\top \mathbf{k}(\mathbf{x}_h^-), \\ G_2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}) = \sum_{j=1}^{q^+} \widetilde{\beta}_j \alpha_j^+ [b - b^* + \boldsymbol{v}_+^\top g^+(\mathbf{x}_j^+)] \\ \qquad\qquad - \sum_{h=1}^{q^-} \widetilde{\gamma}_h \alpha_h^- [b + b^* - \boldsymbol{v}_-^\top g^-(\mathbf{x}_h^-)]. \end{cases}$$

Clearly, the degenerated *pTsm*-SVM in Eq. (25) has a very similar structure to the "pure" black-box SVM in Eq. (**??**) with $G_1(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}})$ and $G_2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}})$ individually representing the terms related to the kernel function and to the Lagrangian multipliers. This makes Eq. (25) look like a "standard" SVM so that the corresponding nonlinear QP problem can be easily solved using LibSVM [34].

The incorporation mechanism of priors into the black-box soft-margin SVM shown above combines ideas from Mangasarian et. al. [22] and Xu et. al. [33]. Despite this fact, there are still some encouraging novelty in formulating *pTsm*-SVM. Except that the prior information is not ready but needs to be mined from data, the QP problem is handled according to **Propositions 4** and **5**, and finally takes the form of Eq. (25) which has the same structure as a "pure" black-box SVM and 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., interpretability together with high precision.

V. AN 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 $g^+(\mathbf{x})$ or $g^-(\mathbf{x})$ 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

---

Algorithm 1. *Mining Data-based Linear Prior Information*

Input: $\{(\mathbf{x}_k, y_k)\}_{k=1}^N, \mathbf{x} = (x^{(1)}, \cdots, x^{(n)})^\top \in \mathbb{R}^n, y_k \in \{+1, -1\}$
Output: Data-based prior information

1:   Data normalization;
2:   Set $\Gamma^+ = \{\mathbf{x}_k | y_k = +1, \forall\ k\}$ and $\Gamma^- = \{\mathbf{x}_k | y_k = -1, \forall\ k\}$;
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 $\mathbf{x}_k$ $(k = 1, \cdots, N)$ to the i-j input feature subspace with images denoted by $\text{Img}_{ij}\mathbf{x}_k$;
      $\circ$ Solve the following optimization problem

$$\max_{\phi, c} \quad \text{Card}\left(\{(x^{+(i)}, x^{+(j)}) | x^{+(i)} \cos\phi + x^{+(j)} \sin\phi + c \le 0\}\right)$$

$$\text{s.t.} \quad x^{-(i)} \cos\phi + x^{-(j)} \sin\phi + c > 0, \ \phi \in [0, 2\pi], \ c \in \mathbb{R},$$
$$(x^{+(i)}, x^{+(j)}) \in \text{Img}_{ij}\Gamma^+, \ (x^{-(i)}, x^{-(j)}) \in \text{Img}_{ij}\Gamma^-.$$

      where $\text{Card}(\cdot)$ 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 $(\hat i, \hat j; \hat\phi_{\hat i \hat j}, \hat c_{\hat i \hat j}) = \arg\max_{i,j} \Omega_{(i,j;\hat\phi_{ij}, \hat c_{ij})}$;
5:   Output positive class prior information

$$\{k | x_k^{(\hat i)} \cos\hat\phi_{\hat i \hat j} + x_k^{(\hat j)} \sin\hat\phi_{\hat i \hat j} + \hat c_{\hat i \hat j} \le 0 \implies y_k = +1\};$$

6:   The same way produces negative class prior information.

---

of system. The main thought of mining data-based linear prior information is to map the sampling points in $\mathbb{R}^n$ to the i-j "input feature subspace", 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 $\mathscr{L}_{ij}^+$ be a straight line representing the positive class prior information function $g^+(\mathbf{x}) = \mathbb{0}_{r^+}$, then it requires that all negative class samplings fall above $\mathscr{L}_{ij}^+$ while positive class samplings fall below $\mathscr{L}_{ij}^+$ as many as possible. The same principle could generate negative class prior information $\mathscr{L}_{ij}^-$. We formulate the mining process as **Algorithm 1**.

In **Algorithm 1**, we propose a cardinality maximization problem. The object function

$$\text{Card}\left(\{(x^{+(i)}, x^{+(j)}) | x^{+(i)} \cos\phi + x^{+(j)} \sin\phi + c \le 0\}\right)$$

represents in the i-j input feature space there are as many as possible positive class samples below the learned hyperplane while the constraint $x^{-(i)} \cos\phi + x^{-(j)} \sin\phi + c > 0$ means all negative class samples to be distributed above the hyperplane. Therefore, the solution must classify right as many as possible positive class samples while classify right all negative class samples. In a similar way, the negative class knowledge must classify right as many as possible negative class samples while classify right all positive class samples. Overall the mined linear priors can complement black-box learning algorithms for classification, and can help to solve the original problem.

Seemingly, the optimization in **Algorithm 1** is a NP hard nonconvex/nonlinear/nonsmooth optimization problem. How-

ever, in fact the hyperplane learned by optimizing ϕ and c is simply a normalized hyperplane (that is a unit-norm hyperplane). Theoretically, the hyperplane optimization could be rewritten as a linear program with respect to the normal vector to the line instead of ϕ and c. If doing so, the cardinality maximization problem would become difficult to be solved with respect to the normal vector. The usage of the angle ϕ instead can avoid this issue. In practical solution procedure, the angel range $[0, 2\pi]$ is discretized into 63 points with step by 0.1, and the hyperplane is rotated along every discretized angle to produce the result. We admit that this result is not the optimal one, but a relaxation solution. It naturally can produce better solution by making more thinning discretization on the angle.

Also, it should be noted in **Algorithm 1** that there needs to solve $\frac{n(n-1)}{2}$ optimization problems to find a piece of prior information. The computational complexity of the algorithm is thus $O(63 \times N \times \frac{n(n-1)}{2})$. It is obvious that the algorithm scalability will increase as the the dimension of data, n, becomes larger. However, compared with the SVM model solution process, **Algorithm 1** will not lead to scalability disaster even if for high-dimensional data, since the former has the computational complexity $O(n \times N^2)$ [34] while the size of the data set, N, is usually far larger than n. This means the algorithm can be generalized to handle general classification problems such as text classification and image classification where high dimensional features will be encountered. We exhibit this algorithm through the following example and also report the running time.

Example 1. **Algorithm 1** is applied to Liver disorders[1], a public data set, to exhibit effectiveness, 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 II reports the mined positive class linear prior information along every projected two dimensional plane. It is clear that the straight line

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

in the 1-4 input feature subspace is the expected one. The running time is about 0.30s, which is very small. In the same way, we can obtain negative class prior information in the 3-5 input feature subspace to be

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

Remark 7: 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 8: 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

---

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets

TABLE II
MINED POSITIVE CLASS LINEAR PRIORS FOR LIVER DISORDERS DATA
SET ALONG TWO DIMENSIONAL INPUT FEATURE SPACES

i-j	$\mathscr{L}^{+}_{(i,j;\hat{\phi}_{ij},\hat{c}_{ij})}$	$\Omega_{(i,j;\hat{\phi}_{ij},\hat{c}_{ij})}$
1-2	$-0.9422x^{(1)} + 0.3350x^{(2)} + 0.7347 \le 0$	5
1-3	$0.1700x^{(1)} + 0.9854x^{(3)} - 0.1651 \le 0$	10
1-4	$0.6347x^{(1)} - 0.7728x^{(4)} - 0.0156 \le 0$	**13**
1-5	$0.5403x^{(1)} + 0.8415x^{(5)} - 0.2648 \le 0$	4
1-6	$-0.9900x^{(1)} + 0.1411x^{(6)} + 0.8689 \le 0$	6
2-3	$-0.7259x^{(2)} - 0.6878x^{(3)} + 0.7686 \le 0$	5
2-4	$-0.7910x^{(2)} - 0.6119x^{(4)} + 0.8974 \le 0$	5
2-5	$-0.8968x^{(2)} - 0.4425x^{(5)} + 0.8473 \le 0$	3
2-6	$-0.8011x^{(2)} + 0.5985x^{(6)} + 0.6779 \le 0$	2
3-4	$0.9950x^{(3)} + 0.0998x^{(4)} - 0.0634 \le 0$	6
3-5	$0.9801x^{(3)} + 0.1987x^{(5)} - 0.0510 \le 0$	4
3-6	$-0.5048x^{(3)} + 0.8632x^{(6)} + 0.1437 \le 0$	6
4-5	$-0.8011x^{(4)} + 0.5985x^{(5)} + 0.3473 \le 0$	7
4-6	$-0.5885x^{(4)} + 0.8085x^{(6)} + 0.2136 \le 0$	9
5-6	$-0.9422x^{(5)} + 0.3350x^{(6)} + 0.4111 \le 0$	5

TABLE III
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

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 9: 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.

VI. NUMERICAL EXPERIMENTS

In this section, the degenerated *pTsm*-SVM in Eq. (25) is used to model 8 benchmark data sets (all available at the same website as given in *Example 1*): 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 III. For these 8 benchmark examples, two-class classification problems are addressed. The priors for every data set are linear and acquired through the proposed mining algorithm in Section V. For the kernel function in Eq. (25), the Gaussian radial basis kernel defined by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma^2) \quad (26)$$

is used. 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. (25). The parameters training is made through grid search together with five-fold crossing validation for the purpose of reducing over-fitting phenomenon. All experiments are carried out in Matlab7.0 environment running on a desktop PC with a 1.80GHz AMD Athlon (tm) Processor and a 4.00 GB memory.

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 IV are the results, where the running time for every data set is also reported. As can be seen, the longest time is 3.61s for Sonar data set which further indicates that **Algorithm 1** has large potential to be generalized to handle general classification problems. Note that the inclusion of priors may bring a great deal of equations into the optimization model, especially when the meshes of points are high. It is possible that these equations are either redundant or causing overfitting in machine learning. To circumvent this issue, we only consider those prior constraints on data points. As an example, there are 13 equations, i.e., $q^+ = 13$, as positive class priors included for Liver data set. After substituting these priors into Eq. (25), respectively, there are $(\nu, \sigma, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}})$ 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 $[\nu_{\min}, \nu_{\max}]$ as the searching range for finding it with $\nu_{\min} = 0.1$ and $\nu_{\max}$ depending on the specific example, calculated by $\nu_{\max} = \frac{2\min(N^+, N^-)}{N}$ [38] where N^+ and N^- represent the number of positive samplings and negative ones, respectively. As for the other three parameters, the searching ranges are $\{2^{-3}, 2^{-2}, \cdots, 2^5, 2^6\}$ for σ, and $\frac{1}{N}$ times of some points in $[0, 1]$ for the components of $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$. We consult the penalty factor of slack variables ξ_i ($i = 1, \cdots, N$) in Eq. (1) to set the searching range for $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$ like so.

After finishing learning these parameters through grid search and five-fold crossing 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

TABLE IV
ACQUIRED PRIOR INFORMATION AND CLASSIFICATION RESULTS OF BENCHMARK DATA SETS

Name of Data Set	Linear Prior Information	**Algorithm 1** time cost (s)	ATA[†] (%) with Priors	ATA (%) without Priors	p-value of the t-test
Australian	$g^+(\mathbf{x}) = -0.3037x^{(5)} - 0.9516x^{(14)} + 0.3085 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = 0.7087x^{(8)} - 0.7055x^{(10)} + 0.0176 \leq 0 \Longrightarrow y = -1$	0.54	85.22 ± 2.70	82.54 ± 6.40	0.03
Breast cancer	$g^+(\mathbf{x}) = 0.8855x^{(1)} - 0.4646x^{(9)} + 0.2890 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = 0.9211x^{(1)} + 0.3894x^{(8)} - 0.0746 \leq 0 \Longrightarrow y = -1$	0.30	97.34 ± 0.53	96.85 ± 0.58	0.00
Diabetes	$g^+(\mathbf{x}) = -0.9900x^{(2)} + 0.1411x^{(7)} + 0.9467 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = 0.3624x^{(2)} + 0.9320x^{(8)} - 0.2004 \leq 0 \Longrightarrow y = -1$	0.30	76.67 ± 1.66	75.92 ± 1.26	0.04
German	$g^+(\mathbf{x}) = 0.6347x^{(2)} - 0.7728x^{(4)} + 0.3679 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = 0.0875x^{(4)} - 0.9962x^{(17)} + 0.9812 \leq 0 \Longrightarrow y = -1$	0.89	75.83 ± 2.44	74.83 ± 2.23	0.05
Heart	$g^+(\mathbf{x}) = -0.3739x^{(10)} + 0.9275x^{(8)} - 0.0547 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = 0.7756x^{(5)} - 0.6317x^{(8)} + 0.4139 \leq 0 \Longrightarrow y = -1$	0.31	84.25 ± 3.74	83.50 ± 4.78	0.26
Ionosphere	$g^+(\mathbf{x}) = -0.6663x^{(5)} + 0.7457x^{(27)} + 0.3465 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = x^{(5)} - 0.5210 \leq 0 \Longrightarrow y = -1$	3.40	93.96 ± 2.23	93.49 ± 1.69	0.21
Liver disorders	$g^+(\mathbf{x}) = 0.6347x^{(1)} - 0.7728x^{(4)} - 0.0156 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = -0.1288x^{(3)} + 0.9917x^{(5)} + 0.0025 \leq 0 \Longrightarrow y = -1$	0.30	73.10 ± 3.70	70.80 ± 3.52	0.02
Sonar	$g^+(\mathbf{x}) = 0.9553x^{(13)} + 0.2955x^{(20)} - 0.2426 \leq 0 \Longrightarrow y = +1$ $g^-(\mathbf{x}) = -0.5748x^{(11)} - 0.8183x^{(27)} + 0.9725 \leq 0 \Longrightarrow y = -1$	3.61	85.72 ± 3.59	83.02 ± 4.55	0.04

[†] represents Average Testing Accuracy.

example, and calculated their average testing accuracy values and the corresponding standard deviations, as shown in Table IV. 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 IV 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 IV. 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 t-test results (p-values) are also reported in Table IV, where the p-value represents the probability in which 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 interpretability 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 interpretable 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

TABLE V
CLASSIFICATION RESULTS ON THE BENCHMARK DATA SETS WITH OTHER MODELS

Name of Data Set	degenerated *pTsm*-SVM	Soft-margin MKL [33]§	$\ell_2$trStMKL [40]†
Australian	85.22 ± 2.70	86.23 ± 1.94	$\varnothing$
Diabetes	76.67 ± 1.66	76.35 ± 2.79	$\varnothing$
German	75.83 ± 2.44	$\varnothing$	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	$\varnothing$	66.20 ± 2.40
Sonar	85.72 ± 3.59	$\varnothing$	83.30 ± 2.60

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

TABLE VI
AVERAGE PROCESSING TIME FOR THE BENCHMARK DATA SETS

Name of Data Set	APT* (s) with Priors	APT (s) without Priors
Australian	728.9	28.4
Breast cancer	220.8	15.5
Diabetes	860.8	21.3
German	1708.5	70.6
Heart	15.1	3.9
Ionosphere	927.7	6.8
Liver disorders	116.6	4.7
Sonar	58.7	4.2

* APT represents Average Processing Time.

the possibility of slight loss in precision for these two data sets, the degenerated *pTsm*-SVM is still a good alternative due to its interpretability 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 V. The model in [33] is a soft-margin multiple kernels SVM while the model in [40] is multiple kernels SVM integrating radius information. These two kinds of models were asserted to be able to outperform other similar models, like MKL, ℓ_pMKL, etc. As can be seen from Table V, 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.

It should be noted that the incorporation of priors into the soft-margin SVM may result in the processing time for the degenerated *pTsm*-SVM to increase greatly, including priors mining and soft-margin SVM training time. There are two additional parameter vectors $\widetilde{\beta}$ and $\widetilde{\gamma}$ in the degenerated *pTsm*-SVM that need to be trained. Table VI reports the average processing time (APT) of the model with and without priors incorporation for the 8 benchmark data sets. As expected just now, the APT for the degenerated *pTsm*-SVM is far beyond

that for the soft-margin SVM, especially in the case of the Ionosphere data set, 927.7s vs. 6.8s. This terrible phenomenon seems to suggest the degenerated *pTsm*-SVM is much poorer than the soft-margin SVM. Actually, as far as these benchmark data sets are concerned, the processing time of the model may be not quite important, such as for the Breast cancer/Heart data sets, the most important should be to diagnose right the cancer/heart disease and also to give the etiology, but not the diagnostic time. Therefore, the reported APT of every benchmark data set with priors in Table VI, despite being very high, is still acceptable from the viewpoint of practical applications. Certainly, there are rather large rooms to reduce the processing time of the degenerated *pTsm*-SVM so that this model looks also efficient. Since the additional processing time mainly comes from training $\widetilde{\beta}$ and $\widetilde{\gamma}$ in Eq. (25), a solution of raising efficiency of the model is to derive their optimal values from theoretical analysis instead of grid search. We put this issue as one of the main points of future research.

In summary, from the viewpoint of model precision, the degenerated *pTsm*-SVM is not sure to be 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 interpretability 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 interpretability enhancement of the black-box soft-margin SVM. Frankly speaking, it is not our original intension to expect the degenerated *pTsm*-SVM better than the competing state-of-the-arts in precision when used for the benchmark data sets. The main contribution of this paper is to provide a way for adding interpretability of black-box models, and the models performance comparison should be made between the interpretable model and the corresponding black-box model. It is interesting to observe the effect if the mined linear priors, maybe in the form of logical implications as given in Eqs. (7) and (14), are incorporated into other black-box models, such as neural network, multiple kernels learning [39], [40], 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 $\widetilde{\beta}$ and $\widetilde{\gamma}$ in Eq. (25), based on theoretical analysis. The effort towards this target is on the way.

VII. CONCLUSIONS AND POINTS OF 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 interpretability 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 common software packages can be directly utilized to find solutions.

Despite the good performance exhibited by the degenerated *pTsm*-SVM, there are still great rooms for the model to be improved. The most urgent task maybe focuses on mining priors as reliable as possible from the data set, such as including more inputs to model linear priors or directly developing nonlinear priors. Moreover, in the case of the linear priors, the formulation of the optimization problem should consider the entire regions of feature space defined by the prior knowledge constraints, but no longer depends on $\mathbf{x}^+$ and $\mathbf{x}^-$ as individual data points or meshes of points. After all, the current improvements in data sets are very small when prior knowledge is added while a lot of previous results with prior knowledge report significant accuracy improvements [22], [36] of anywhere from 15% to 50%. In addition, more concerns should be thrown towards reducing the training time of the interpretable 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 interpretable model has the same number of parameters to be trained with the corresponding black-box model. Certainly, it is interesting to achieve the theoretical support that the degenerated *pTsm*-SVM is able to converge the true one, i.e., to the corresponding Bayes model. The estimation of the convergence rate is also an important investigation point.

Finally, it should be pointed out that the prior information learned is irrelevant with SVM. Even if the problem described in Eq. (17) is derived by imitating the mechanism of SVM, it is actually not specialized for improving the performance of SVM. Those prior information corrected from Eq. (17) can also be used in other machine learning algorithms such as naive Bayes or deep neural network. Therefore, it is interesting to investigate the incorporation of the mined priors into other machine learning algorithms for interpretability enhancement.

REFERENCES

[1] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, "ERNN: A biologically inspired feedforward neural network to discriminate emotion from EEG signal," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 609-620, Mar. 2014.

[2] D. Grumiller, R. McNees, and S. Zonetti, "Black holes in the conical ensemble," *Phys. Rev. D*, vol. 86, no. 12, pp. 124043, Dec. 2012.

[3] C. H. Gao, L. Jian, and S. H. Luo, "Modeling of the thermal state change of blast furnace hearth with support vector machine," *IEEE Trans. Ind. Electron.*, vol. 59, no. 2, pp. 1134-1145, Feb. 2012.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," *ACM Conf. Knowledge Discovery and Data Mining*, pp. 1135-1144, Aug. 2016.

[5] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector regression," *Mach. Learn.*, vol. 70, no. 1, pp. 89-118, Jan. 2008.

[6] Y. J. Qu and B. G. Hu, "Generalized contraint neural network regression model subject to linear priors," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2447-2459, Dec. 2011.

[7] R. V. Borges, A. d'A. Garcez, and L. C. Lamb, "Learning and representing temporal knowledge in recurrent networks," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2409-2421, Dec. 2011.

[8] D. Martens, B. Baesens, and T. V. Gestel, "Decompositional rule extraction from support vector machines by active learning," *IEEE Trans. Knowledge and Data Eng.*, vol. 21, no. 2, pp. 178-191, Feb. 2009.

[9] J. Chorowski and J. M. Zurada, "Extracting rules from neural networks as decision diagrams," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2435-2446, Dec. 2011.

[10] T. Q. Huynh and J. A. Reggia, "Guiding hidden layer representations for improved rule extraction from neural networks," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 264-275, Feb. 2011.

[11] N. Barakat and A. P. Bradley, "Rule extraction from support vector machines: A review," *Neurocomputing*, vol. 74, no. 1-3, pp. 178-190, Dec. 2010.

[12] D. Wolpert, W. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67-82, April 1997.

[13] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector machines for classification: A review," *Neurocomputing*, vol. 71, no. 7-9, pp. 1578-1594, Mar. 2008.

[14] P. Niyogi, F. Girosi, and T. Poggio, "Incorporating prior information in machine learning by creating virtual examples," *Proc. IEEE*, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.

[15] B. G. Hu, Y. Wang, S. H. Yang, and H. B. Qu, "How to add transparency to artificial neural networks," *Pattern Recognit. Artif. Intell.*, vol. 20, no. 1, pp. 72-84, Feb. 2007.

[16] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[17] S. Chen, A. Wolfgang, C. Harris, and L. Hanzo, "Symmetric RBF classifier for nonlinear detection in multiple-antenna-aided systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 5, pp. 737-745, May 2008.

[18] Y. S. Abu-Mostafa, "Learning from hints in neural networks," *J. Complexity*, vol. 6, no. 2, pp. 192-198, June 1990.

[19] H. Daniels and M. Velikova, "Monotone and partially monotone neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 6, pp. 906-917, June 2010.

[20] O. L. Mangasarian, J. W. Shavlik, and E. W. Wild, "Knowledge-based kernel approximation," *J. Mach. Learn. Res.*, vol. 5, pp. 1127-1141, Sep. 2004.

[21] O. L. Mangasarian and E. W. Wild, "Nonlinear knowledge in kernel approximation," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 300-306, Jan. 2007.

[22] O. L. Mangasarian and E. W. Wild, "Nonlinear knowledge-based classification," *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1826-1832, Oct. 2008.

[23] L. Aguirre, R. Lopes, G. Amaral, and C. Letellier, "Constraining the topology of neural networks to ensure dynamics with symmetry properties," *Phys. Rev. E*, vol. 69, no. 2, pp. 026701, 2004.

[24] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481-1497, Sep. 1990.

[25] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[26] G. G. Towell and J. W. Shavlik, "Knowledge-based artificial neural networks," *Artificial Intell.*, vol. 70, no. 1/2, pp. 119-165, Oct. 1994.

[27] A. Minin, M. Velikova, B. Lang, and H. Daniels, "Comparison of universal approximators incorporating partial monotonicity by structure," *Neural Netw.*, vol. 23, no. 4, pp. 471-475, May 2010.

[28] H. Daniels and M. Velikova, "Monotone and Partially Monotone Neural Networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 6, pp. 906-917, June 2010.

[29] H. Velikova, H. Daniels, and A. Feelders, "Mixtures of monotone networks for prediction," *Int. J. Comput. Intell.*, vol. 3, no. 3, pp. 204-214, 2006.

[30] C. Dugas, Y. Bengio, F. Belisle, C. Nadeau, and R. Garcia, "Incorporating functional knowledge in neural networks," *J. Mach. Learn. Res.*, vol. 10, no. 1, pp. 1239-1262, June 2009.

[31] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing, "Deep neural networks with massive learned knowledge," *Conf. Empirical Methods in Natural Language Processing*, pp. 1670-1679, 2016.

[32] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[33] X. X. Xu, I. W. Tsang, and D. Xu, "Soft mMargin multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749-761, May 2013.

[34] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," 2001. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm.

[35] M. A. Goberna and M. A. Lopez, *Linear Semi-infinite Optimization*, John Wiley, New York, 1998.

[36] G. Fung, O. L. Mangasarian, and J. W. Shavlik. "Knowledge-based support vector classifiers," *Advances in Neural Information Processing Systems*, vol. 15, pp. 521-528, 2003.

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

[38] P. H. Chen, C. J. Lin and B. Schölkopf, "A tutorial on v-support vector machines," *Appl. Stochastic Models Bus. Ind.*, vol. 21, no. 2, pp. 111-136, Mar.-April 2005.

[39] M. Hu, Y. Chen, and J. T. Kwok, "Building sparse multiple-kernel SVM classifiers," *IEEE Trans. Neural Netw.*, vol. 20, no. 5, pp. 827-839, May 2009.

[40] X. Liu, L. Wang, J. Yin, E. Zhu, and J. Zhang, "An efficient approach to integrating radius information into multiple kernel learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 557-569, April 2013.

Shaohan Chen received the B.S. degrees in Mathematics and Applied Mathematics from Jimei University, China, in 2014. She is currently working towards the Ph.D. degree in operational research and cybernetics at Zhejiang University.

Her research interests are in the areas of Machine learning and Transparency of black-box modeling techniques.

Chuanhou Gao (M'09 SM'12) received the B.Sc. degrees in Chemical Engineering from Zhejiang University of Technology, China, in 1998, and the Ph.D. degrees in Operational Research and Cybernetics from Zhejiang University, China, in 2004. From June 2004 until May 2006, he was a Postdoctor in the Department of Control Science and Engineering at Zhejiang University.

Since June 2006, he has joined the Department of Mathematics at Zhejiang University, where he is currently a Professor. He was a visiting scholar at Carnegie Mellon University from Oct. 2011 to Oct. 2012. His research interests are in the areas of data-driven modeling, control and optimization, chemical reaction network theory and thermodynamic process control. He is a guest editor of IEEE Transactions on Industrial Informatics, ISIJ International and Journal of Applied Mathematics, and an editor of Metallurgical Industry Automation from May 2013.

Ping Zhang received the B.S. degrees in applied information and computation science from China Ji Liang Univercity, China, in 2012. And the M.S. degrees in operational research cybernetics from Zhejiang University, China, in 2014.

Since July 2014, he has joined the ZTE Co. , Ltd. , Shanghai, China. His research interests are in the areas of Machine learning and Transparency of black-box modeling techniques.