R-SVM+: Robust Learning with Privileged Information

Xue Li[1,2], **Bo Du**[*1], **Chang Xu**[3], **Yipeng Zhang**[1], **Lefei Zhang**[1], **Dacheng Tao**[3]
[1] School of Computer Science, Wuhan University, China
[2] LIESMARS, Wuhan University, China
[3] UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Australia
{lixue93, remoteking}@whu.edu.cn, c.xu@sydney.edu.au,
{zyp91, zhanglefei@whu}@whu.edu.cn, dacheng.tao@sydney.edu.au

Abstract

In practice, the circumstance that training and test data are clean is not always satisfied. The performance of existing methods in the learning using privileged information (LUPI) paradigm may be seriously challenged, due to the lack of clear strategies to address potential noises in the data. This paper proposes a novel Robust SVM+ (R-SVM+) algorithm based on a rigorous theoretical analysis. Under the SVM+ framework in the LUPI paradigm, we study the lower bound of perturbations of both example feature data and privileged feature data, which will mislead the model to make wrong decisions. By maximizing the lower bound, tolerance of the learned model over perturbations will be increased. Accordingly, a novel regularization function is introduced to upgrade a variant form of SVM+. The objective function of R-SVM+ is transformed into a quadratic programming problem, which can be efficiently optimized using off-the-shelf solvers. Experiments on real-world datasets demonstrate the necessity of studying robust SVM+ and the effectiveness of the proposed algorithm.

1 Introduction

Great advances in machine learning have been inspired by the deeper investigation into the learning process of human beings [Shen *et al.*, 2014; Wang *et al.*, 2018]. Conventional supervised methods utilize the priori knowledge to help us understand the world. Based on a set of examples and their corresponding labels, traditional supervised methods can train a classification model, and then use it to classify unknown test examples. However, in practice, there often exists some auxiliary information associated with an example except for its label. This auxiliary information can be widely found in human teaching and learning process. For example, the teacher plays an important role to provide students with helpful comments, comparisons, and explanations to improve students' performance. Inspired by this fact, Vapnik and Vashist [Vapnik and Vashist, 2009] introduced the paradigm of learning using privileged information (LUPI) that focuses on improving the learning with the auxiliary information which is supplied by a teacher about examples at the training stage. Since this auxiliary information will not be available at the test stage, it is referred to as privileged information.

As one of the most popular classifiers, support vector machine (SVM) was first upgraded in the paradigm of learning using privileged information (LUPI) [Vapnik and Vashist, 2009], and the new method is called SVM+. The main idea of SVM+ is to define a linear or nonlinear correcting (slack) function in the privileged feature space to estimate the slack variables in the standard SVM method using privileged information. Recently, an increasing attention has been attracted on the LUPI paradigm [Vapnik and Izmailov, 2015; Motiian *et al.*, 2016; Zhou *et al.*, 2016; Yang *et al.*, 2017], and some SVM+-based algorithms have been proposed and applied for various applications. Beyond L-2 SVM, privileged information is also introduced into an L-1 regularized SVM to reduce time consumption on tuning model parameters [Niu *et al.*, 2012]. Considering privileged label information in the multi-label learning problems, a privileged multi-label learning (PrML) method explores and exploits the connections between different examples' labels and is extended into domain adaptation [You *et al.*, 2017]. Moreover, there are also multi-task multi-class SVM+ [Ji *et al.*, 2012], structural SVM+ [Feyereisl *et al.*, 2014], and the rank transfer method [Sharmanska *et al.*, 2013]. Various optimization techniques to solve SVM+ have been studied recently, such as MAT-SVM+, CVX-SVM+, and L2-loss SVM+ [Li *et al.*, 2016]. In [Li *et al.*, 2016], two new algorithms are proposed to efficiently solve linear SVM+ and kernel SVM+ which uses the L2-loss based on the ρ-SVM formulation, respectively.

These methods have largely advanced the developments on LUPI. However, their successes are usually achieved in the circumstance that training and test data are deemed to be clean and the teacher always makes correct judgement. In

*Corresponding author: Bo Du.

practice, we can well design the training set under our demands, but it is difficult and even impossible to tell what test data will be. Existing methods lack clear strategies to address potential noises in the data, and thus their practical performances will be seriously deteriorated. In addition, existing methods in LUPI used to consider that teacher's comments are always accurate. But if there exist noises in the data, teacher may not be guaranteed to make correct judgements any more, which will then influence the student performance as a result.

In this paper, we derive a novel Robust SVM+ (R-SVM+) algorithm based on a rigorous theoretical analysis. Considering perturbations over both example feature data and privileged feature data, we study the lower bound of these perturbations that will mislead the model to make wrong judgements. A novel regularization function adapted from this lower bound is introduced to upgrade a variant form of SVM+. In this way, the capability of the learned model to tolerate perturbations over the data will be enhanced - that is to say the robustness of the model will be strengthened. The objective function of R-SVM+ is transformed into a quadratic programming problem, which can be efficiently optimized using off-the-shelf solvers. Experimental results demonstrate the necessity of researching robust SVM+ and the effectiveness of the proposed algorithm.

2 Preliminary of LUPI

The LUPI paradigm considers a set of training examples where privileged information is additionally supplied,

$$(\boldsymbol{x_1}, \boldsymbol{x_1^*}, y_1), (\boldsymbol{x_2}, \boldsymbol{x_2^*}, y_2), \ldots, (\boldsymbol{x_n}, \boldsymbol{x_n^*}, y_n),$$

where $\boldsymbol{x_i} \in \mathbb{R}^d$ and $\boldsymbol{x_i^*} \in \mathbb{R}^{d^*}$ are the i-th example feature (EF) vector and its corresponding privileged feature (PF) vector, $y_i \in \{+1, -1\}$ is the ground-truth label of the i-th training example pair $(\boldsymbol{x_i}, \boldsymbol{x_i^*})$, and n is the number of training example pairs.

The first approach proposed in the LUPI paradigm is called SVM+ [Vapnik and Vashist, 2009], which tries to measure the misclassification loss of training example with a correcting function learned from privileged information. The objective function of SVM+ can be formulated as follows:

$$\begin{aligned} \min_{\substack{\boldsymbol{w},\boldsymbol{w^*},\\ b,b^*}} \ & \frac{1}{2}(\langle \boldsymbol{w}, \boldsymbol{w}\rangle + \rho\langle \boldsymbol{w^*}, \boldsymbol{w^*}\rangle) + C\sum_{i=1}^{n}[\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*] \\ s.t. \ & y_i[\langle \boldsymbol{w}, \phi(\boldsymbol{x_i})\rangle + b] \geq 1 - [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*], \\ & \langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^* \geq 0, \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{w}$ and $\boldsymbol{w^*}$ are the weight vectors, b and b^* are the bias terms, C is the a non-negative parameter which balances the loss term and the regularizer, the term $\frac{\rho}{2}\langle \boldsymbol{w^*}, \boldsymbol{w^*}\rangle$ in Eq. (1) aims to restrict the capacity of the correcting function space, $\rho > 0$ is the trade-off parameter, the functions $\phi(\cdot)$ and $\psi(\cdot)$ are two feature mappings induced by the kernels on example features and privileged features, respectively, and $\langle \boldsymbol{a}, \boldsymbol{e}\rangle$ denotes the inner product between two vectors $\boldsymbol{a}$ and $\boldsymbol{e}$.

By introducing Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$, $i = 1, \ldots, n$, we can arrive at the dual form of SVM+,

$$\begin{aligned} \max_{\alpha,\beta} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ & - \frac{1}{2\rho}\sum_{i,j=1}^{n}(\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C)k^*(\boldsymbol{x_i^*}, \boldsymbol{x_j^*}), \end{aligned} \tag{2}$$

subject to constraints $\sum_{i=1}^{n}(\alpha_i + \beta_i - C) = 0$, $\sum_{i=1}^{n}\alpha_i y_i = 0$, $\alpha_i \geq 0$, and $\beta_i \geq 0$, $i = 1, \ldots, n$, and where $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j})\rangle$ and $k^*(\boldsymbol{x_i^*}, \boldsymbol{x_j^*}) = \langle \psi(\boldsymbol{x_i^*}), \psi(\boldsymbol{x_j^*})\rangle$ are kernels in example feature and privileged feature spaces, respectively. After solving the dual optimization problem, two weight vectors can be reconstructed as $\boldsymbol{w} = \sum_{i=1}^{n}\alpha_i y_i \phi(\boldsymbol{x_i})$ and $\boldsymbol{w^*} = \frac{1}{\rho}\sum_{i=1}^{n}(\alpha_i + \beta_i - C)\psi(\boldsymbol{x_i^*})$.

3 Robust SVM+

According to the constraint of Eq. (1), we can define two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x^*})$, where $f(\boldsymbol{x}) = 1 - y[\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b]$ denotes a hinge loss of the decision function $\hbar = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b$ on the example $(\boldsymbol{x}, y)$ and $g(\boldsymbol{x^*}) = \langle \boldsymbol{w^*}, \psi(\boldsymbol{x^*})\rangle + b^*$ denotes a loss of the correcting function on the privileged data $\boldsymbol{x^*}$. And $f(\boldsymbol{x})$ and $g(\boldsymbol{x^*})$ under the framework of SVM+ are

$$\begin{aligned} f(\boldsymbol{x}) &= 1 - y[\sum_{i=1}^{n}\alpha_i y_i k(\boldsymbol{x_i}, \boldsymbol{x}) + b], \\ g(\boldsymbol{x^*}) &= \frac{1}{\rho}\sum_{i=1}^{n}(\alpha_i + \beta_i - C)k^*(\boldsymbol{x_i^*}, \boldsymbol{x^*}) + b^*. \end{aligned} \tag{3}$$

In the SVM+ method, the constraint $y_i[\langle \boldsymbol{w}, \phi(\boldsymbol{x_i})\rangle + b] \geq 1 - [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*]$ in Eq. (1) can be written as Eq. (4). The inequality in Eq. (4) will be satisfied if $\boldsymbol{x}$ and $\boldsymbol{x^*}$ are good enough for model training and a small loss $f(\boldsymbol{x})$ in the decision space can be achieved. It is the basic assumption of LUPI that if a small loss in the correcting space can be obtained, then a small loss in the decision space should also be achieved [Pechyony and Vapnik, 2010],

$$f(\boldsymbol{x}) \leq g(\boldsymbol{x^*}). \tag{4}$$

However, in practice there may be some noises over $\boldsymbol{x}$ and $\boldsymbol{x^*}$ sometimes. If perturbations caused by noises over examples are large enough, the inequality will not hold, which will then influence the correct decisions of the model on examples. In response to this circumstance, we propose to learn a Robust SVM+ (R-SVM+) algorithm, which has a stronger capability to tolerate perturbations over the data caused by noises. We assume there are some perturbations $\boldsymbol{\tau_x} \in \mathbb{R}^d$ and $\boldsymbol{\tau_{x^*}} \in \mathbb{R}^{d^*}$ over the ideal observations $\boldsymbol{x}$ and $\boldsymbol{x^*}$, respectively, i.e., $\boldsymbol{x} + \boldsymbol{\tau_x}$ and $\boldsymbol{x^*} + \boldsymbol{\tau_{x^*}}$, which are large enough to make the following inequality satisfied,

$$f(\boldsymbol{x} + \boldsymbol{\tau_x}) > g(\boldsymbol{x^*} + \boldsymbol{\tau_{x^*}}). \tag{5}$$

Our purpose is to study what properties $\boldsymbol{\tau_x}$ and $\boldsymbol{\tau_{x^*}}$ should have if Eq. (5) holds. Therefore, in the following, we proceed to show that these perturbations $\boldsymbol{\tau_x}$ and $\boldsymbol{\tau_{x^*}}$ actually have the

lower bounds based on a rigorous theoretical analysis. According to the theorem of calculus, we have

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{\tau_x}) &= f(\boldsymbol{x}) + \int_0^1 \langle \nabla f(\boldsymbol{x}+t\boldsymbol{\tau_x}), \boldsymbol{\tau_x}\rangle dt, \\ g(\boldsymbol{x^*}+\boldsymbol{\tau_{x^*}}) &= g(\boldsymbol{x^*}) + \int_0^1 \langle \nabla g(\boldsymbol{x^*}+t\boldsymbol{\tau_{x^*}}), \boldsymbol{\tau_{x^*}}\rangle dt. \end{aligned} \tag{6}$$

If the perturbations are serious enough, according to Eq. (5) and Eq. (6), the following inequality is satisfied,

$$\begin{aligned} &0 \le g(\boldsymbol{x^*}) - f(\boldsymbol{x}) \\ &< \int_0^1 \langle \nabla f(\boldsymbol{x}+t\boldsymbol{\tau_x}), \boldsymbol{\tau_x}\rangle dt - \int_0^1 \langle \nabla g(\boldsymbol{x^*}+t\boldsymbol{\tau_{x^*}}), \boldsymbol{\tau_{x^*}}\rangle dt \\ &= \int_0^1 [\nabla f(\boldsymbol{x}+t\boldsymbol{\tau_x}); -\nabla g(\boldsymbol{x^*}+t\boldsymbol{\tau_{x^*}})]^{\mathrm{T}} \cdot [\boldsymbol{\tau_x}; \boldsymbol{\tau_{x^*}}] dt \\ &\le \|\boldsymbol{\tau}\|_p \int_0^1 \|\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})\|_q dt, \end{aligned}$$

where $\boldsymbol{\tau} = [\boldsymbol{\tau_x}; \boldsymbol{\tau_{x^*}}] \in \mathbb{R}^{d+d^*}$, $\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}}) = [\nabla f(\boldsymbol{x} + t\boldsymbol{\tau_x}); -\nabla g(\boldsymbol{x^*}+t\boldsymbol{\tau_{x^*}})] \in \mathbb{R}^{d+d^*}$ and we have applied Hölder inequality in the last step that the q-norm is dual to the p-norm, where p and q satisfy $\frac{1}{p} + \frac{1}{q} = 1$.

Hence, given $g(\boldsymbol{x^*}) \ge f(\boldsymbol{x})$, we have the minimal perturbations $\boldsymbol{\tau}$ that is required to reverse the decision of the SVM+ classifier,

$$\|\boldsymbol{\tau}\|_p > \frac{g(\boldsymbol{x^*}) - f(\boldsymbol{x})}{\int_0^1 \|\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})\|_q dt}. \tag{7}$$

Eq. (7) indicates the lower bound over the perturbations to bring in the undesirable error in Eq. (5). In order to obtain a more robust classifier, we consider maximizing the lower bound (i.e., the right of Eq. (7)). As a result, the new model will have more tolerances over the perturbations, and will thus be more robust. And this is the main idea of our proposed R-SVM+ algorithm. That is to say, we wants $\int_0^1 \|\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})\|_q dt$ to be small as well as $g(\boldsymbol{x^*}) - f(\boldsymbol{x})$ to be large. Therefore, the new objective function of our proposed R-SVM+ have two components: 1) For $f(\boldsymbol{x}) \le g(\boldsymbol{x^*}) + \epsilon$, ϵ should be minimized; 2) minimizing the value of function $\Psi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}}) = \int_0^1 \|\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})\|_q dt$.

3.1 Minimizing ϵ

First, we consider minimizing ϵ under the framework of SVM+. We replace the constraints $f(\boldsymbol{x_i}) \le g(\boldsymbol{x_i^*}), i = 1, \ldots, n$ in the original SVM+ objective function with the constraints $f(\boldsymbol{x_i}) \le g(\boldsymbol{x_i^*}) + \epsilon_i$, where $\epsilon_i \ge 0, i = 1, \ldots, n$. Then the optimization problem of this variant form of the SVM+ problem can be transformed into,

$$\begin{aligned} \min_{\boldsymbol{w}, \boldsymbol{w^*}, b, b^*, \epsilon} \quad & \frac{1}{2}(\langle \boldsymbol{w}, \boldsymbol{w}\rangle + \rho\langle \boldsymbol{w^*}, \boldsymbol{w^*}\rangle) \\ & + C\sum_{i=1}^n [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i}^*)\rangle + b^*] + \sigma \sum_{i=1}^n \epsilon_i \\ s.t. \quad & y_i[\langle \boldsymbol{w}, \phi(\boldsymbol{x_i})\rangle + b] \ge 1 - [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*] - \epsilon_i, \\ & \langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^* \ge 0, \\ & \epsilon_i \ge 0, \quad i = 1, \ldots, n, \end{aligned} \tag{8}$$

where $\sigma > 0$ is a tradeoff parameter. Similar variant of Eq. (8) has been discussed in [Vapnik and Vashist, 2009] as well, while in this paper we try to make the value $g(\boldsymbol{x^*}) - f(\boldsymbol{x})$ large to enhance the robustness according to Eq. (7). In order to minimize ϵ, the value of σ should be relatively larger compared with C, to reinforce the effect of the smooth function term $\sum_{i=1}^n \epsilon_i$ on the solution. For simplicity, we let $\sigma = \sigma' C$, where $\sigma' > 1$. By introducing Lagrange multipliers $\alpha_i \ge 0$, $\beta_i \ge 0$ and $\eta_i \ge 0$, where $i = 1, \ldots, n$, the Lagrangian is constructed as

$$\begin{aligned} &L(\boldsymbol{w}, \boldsymbol{w^*}, b, b^*, \epsilon, \alpha, \beta, \eta) \\ &= \frac{1}{2}(\langle \boldsymbol{w}, \boldsymbol{w}\rangle + \rho\langle \boldsymbol{w^*}, \boldsymbol{w^*}\rangle) + \sum_{i=1}^n (\sigma' C - \alpha_i - \eta_i)\epsilon_i \\ &+ C\sum_{i=1}^n [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*] - \sum_{i=1}^n \beta_i [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*] \\ &- \sum_{i=1}^n \alpha_i \left\{ y_i[\langle \boldsymbol{w}, \phi(\boldsymbol{x_i})\rangle + b] - 1 + [\langle \boldsymbol{w^*}, \psi(\boldsymbol{x_i^*})\rangle + b^*] \right\}. \end{aligned}$$

We set the derivatives of the Lagrangian function with respect to $\boldsymbol{w}$, $\boldsymbol{w^*}$, b, b^*, ϵ to zeros, and then the Karush–Kuhn–Tucker (KKT) conditions can be obtained. Accordingly, the dual problem of Eq. (8) can be rewritten as,

$$\begin{aligned} \max_{\alpha, \beta} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ & - \frac{1}{2\rho}\sum_{i,j=1}^n (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) k^*(\boldsymbol{x_i^*}, \boldsymbol{x_j^*}), \end{aligned}$$

subject to constraints $\sum_{i=1}^n (\alpha_i + \beta_i - C) = 0$, $\sum_{i=1}^n \alpha_i y_i = 0$, $0 \le \alpha_i \le \sigma' C$, and $\beta_i \ge 0$, where $i = 1, \ldots, n$. We denote $\boldsymbol{\alpha} \circ \boldsymbol{y}$ as the element-wise product between vectors $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^{\mathrm{T}} \in \mathbb{R}^n$ and $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}} \in \mathbb{R}^n$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]^{\mathrm{T}} \in \mathbb{R}^n$, $\boldsymbol{1} = [1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^n$, and $\boldsymbol{C} = [C, \ldots, C]^{\mathrm{T}} \in \mathbb{R}^n$. The dual problem can be further reformulated as

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \quad & \boldsymbol{1}^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha} \circ \boldsymbol{y})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{\alpha} \circ \boldsymbol{y}) \\ & - \frac{1}{2\rho}(\boldsymbol{\alpha} + \boldsymbol{\beta} - \boldsymbol{C})^{\mathrm{T}} \boldsymbol{K^*} (\boldsymbol{\alpha} + \boldsymbol{\beta} - \boldsymbol{C}), \end{aligned} \tag{9}$$

subject to constraints $\boldsymbol{1}^{\mathrm{T}}(\boldsymbol{\alpha} + \boldsymbol{\beta} - \boldsymbol{C}) = 0$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0$, $0 \le \alpha_i \le \sigma' C$, and $\beta_i \ge 0, i = 1, \ldots, n$. $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix based on example features whose each element being $K_{ij} = k(\boldsymbol{x_i}, \boldsymbol{x_j})$ and $\boldsymbol{K^*}$ is the kernel matrix based on privileged features whose each element being $K_{ij}^* = k^*(\boldsymbol{x_i^*}, \boldsymbol{x_j^*}) \in \mathbb{R}^{n \times n}$.

3.2 Minimizing $\Psi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})$

To minimize $\Psi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}}) = \int_0^1 \|\varpi(t, \boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}})\|_q dt$, we can first define the upper bound of perturbations over some fixed range $\Omega_p(\boldsymbol{x}, \ell) = \{\boldsymbol{z} \in \mathbb{R}^d | \|\boldsymbol{x} - \boldsymbol{z}\|_p \le \ell\}$. In this way, we ensure that the upper bound of $\boldsymbol{\tau_x}$ and $\boldsymbol{\tau_{x^*}}$ is at most ℓ by making assertions of perturbations $\boldsymbol{\tau_x}, \boldsymbol{\tau_{x^*}} \in \Omega_p(\boldsymbol{0}, \ell)$. We

further define $\boldsymbol{z} = \boldsymbol{x} + t\boldsymbol{\tau_x}$ and $\boldsymbol{z^*} = \boldsymbol{x^*} + t\boldsymbol{\tau^*}$, $0 < t \leq 1$. In the following, we discuss the case where $p = q = 2$ for simplicity, and other cases will be studied in our future work. Then the inequality holds

$$\begin{aligned}&\sup_{\substack{\boldsymbol{\tau_x},\boldsymbol{\tau_{x^*}}\\ \in\Omega_2(\mathbf{0},\ell)}} \int_0^1 \|\varpi(t,\boldsymbol{\tau_x},\boldsymbol{\tau_{x^*}})\|_2 dt \leq \max_{\substack{\boldsymbol{\tau_x},\boldsymbol{\tau_{x^*}}\\ \in\Omega_2(\mathbf{0},\ell)}} \|\varpi(t,\boldsymbol{\tau_x},\boldsymbol{\tau_{x^*}})\|_2 \\ &= \max_{\substack{\boldsymbol{z}\in\Omega_2(\boldsymbol{x},\ell),\\ \boldsymbol{z^*}\in\Omega_2(\boldsymbol{x^*},\ell)}} \|[\nabla f(\boldsymbol{z}); -\nabla g(\boldsymbol{z^*})]\|_2.\end{aligned} \tag{10}$$

Thus, our problem can be further transformed into minimizing the value of the upper bound, i.e., the right side of Eq. (10). Naturally, we consider using a surrogate $\Theta(f,g)$ of the quantity of right side of Eq. (10) for regularization. According to Eq. (3), we can obtain $\nabla f(\boldsymbol{x}) = -y\sum_{i=1}^n \alpha_i y_i \nabla_{\boldsymbol{x}} k(\boldsymbol{x_i},\boldsymbol{x})$ and $\nabla g(\boldsymbol{x^*}) = \frac{1}{\rho}\sum_{i=1}^n (\alpha_i + \beta_i - C)\nabla_{\boldsymbol{x^*}} k(\boldsymbol{x_i^*},\boldsymbol{x^*})$. Here we use the Gaussian kernel $k(\boldsymbol{x_i},\boldsymbol{x_j}) = e^{-\gamma\|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}$ and we set $\gamma = 1/D$ in the experiment, where D is the mean of distances among examples in the training set. And a novel regularization function $\Theta(f,g)$ is defined as follows,

$$\begin{aligned}\Theta(f,g) &= \frac{1}{n}\sum_{s=1}^n \|[\nabla f(\boldsymbol{x_s}); -\nabla g(\boldsymbol{x_s^*})]\|_2^2 \\ &= \frac{1}{n}\sum_{s=1}^n \| - y_s \sum_{i=1}^n \alpha_i y_i \nabla_{\boldsymbol{x_s}} k(\boldsymbol{x_i},\boldsymbol{x_s})\|_2^2 \\ &+ \frac{1}{n}\sum_{s=1}^n \| - \frac{1}{\rho}\sum_{i=1}^n (\alpha_i + \beta_i - C)\nabla_{\boldsymbol{x_s^*}} k(\boldsymbol{x_i^*},\boldsymbol{x_s^*})\|_2^2 \\ &= \frac{1}{n}\sum_{s=1}^n \sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j h(\boldsymbol{x_i},\boldsymbol{x_j},\boldsymbol{x_s}) \\ +\frac{1}{n\rho^2}\sum_{s=1}^n \sum_{i,j=1}^n &(\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) h^*(\boldsymbol{x_i^*},\boldsymbol{x_j^*},\boldsymbol{x_s^*}),\end{aligned} \tag{11}$$

where $h(\boldsymbol{x_i},\boldsymbol{x_j},\boldsymbol{x_s}) = \langle \nabla_{\boldsymbol{x_s}} k(\boldsymbol{x_i},\boldsymbol{x_s}), \nabla_{\boldsymbol{x_s}} k(\boldsymbol{x_j},\boldsymbol{x_s})\rangle = 4\gamma^2 \langle \boldsymbol{x_s} - \boldsymbol{x_i}, \boldsymbol{x_s} - \boldsymbol{x_j}\rangle e^{-\gamma\|\boldsymbol{x_i}-\boldsymbol{x_s}\|_2^2} e^{-\gamma\|\boldsymbol{x_j}-\boldsymbol{x_s}\|_2^2}$ and $h^*(\boldsymbol{x_i^*},\boldsymbol{x_j^*},\boldsymbol{x_s^*}) = \langle \nabla_{\boldsymbol{x_s^*}} k(\boldsymbol{x_i^*},\boldsymbol{x_s^*}), \nabla_{\boldsymbol{x_s^*}} k(\boldsymbol{x_j^*},\boldsymbol{x_s^*})\rangle = 4\gamma^2 \langle \boldsymbol{x_s^*} - \boldsymbol{x_i^*}, \boldsymbol{x_s^*} - \boldsymbol{x_j^*}\rangle e^{-\gamma\|\boldsymbol{x_i^*}-\boldsymbol{x_s^*}\|_2^2} e^{-\gamma\|\boldsymbol{x_j^*}-\boldsymbol{x_s^*}\|_2^2}$. And we can define a matrix $\boldsymbol{H_s}$ with each element being $H_{s,ij} = h(\boldsymbol{x_i},\boldsymbol{x_j},\boldsymbol{x_s}) \in \mathbb{R}^{n\times n}$ and a matrix $\boldsymbol{H_s^*}$ with each element being $H^*_{s,ij} = h^*(\boldsymbol{x_i^*},\boldsymbol{x_j^*},\boldsymbol{x_s^*}) \in \mathbb{R}^{n\times n}$. And Eq. (11) can be further reformulated as

$$\begin{aligned}\Theta(f,g) = &\frac{1}{n}\sum_{s=1}^n (\boldsymbol{\alpha}\circ\boldsymbol{y})^{\mathrm{T}} \boldsymbol{H_s} (\boldsymbol{\alpha}\circ\boldsymbol{y}) \\ &+ \frac{1}{n\rho^2}\sum_{s=1}^n (\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C})^{\mathrm{T}} \boldsymbol{H_s^*} (\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}).\end{aligned} \tag{12}$$

3.3 Objective Function of R-SVM+

The objective of the proposed R-SVM+ algorithm aims to solve a maximization problem in Eq. (9) and a minimization problem in Eq. (12) at the same time. Therefore, we arrive at the objective function of R-SVM+ which is a minimization problem,

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad &\frac{1}{2}(\boldsymbol{\alpha}\circ\boldsymbol{y})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{\alpha}\circ\boldsymbol{y}) - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\ &+ \frac{1}{2\rho}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C})^{\mathrm{T}} \boldsymbol{K^*} (\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}) + \lambda\Theta(f,g),\end{aligned} \tag{13}$$

subject to $\mathbf{1}^{\mathrm{T}}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}) = 0$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0$, $0 \leq \alpha_i \leq \sigma' C$, and $\beta_i \geq 0, \quad i = 1,\ldots,n$. And λ is a trade-off parameter to control the effect of the proposed regularization term. Eq. (13) can be reformulated with a simple calculation,

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad &\frac{1}{2}(\boldsymbol{\alpha}\circ\boldsymbol{y})^{\mathrm{T}} (\boldsymbol{K} + \frac{2\lambda}{n}\sum_{s=1}^n \boldsymbol{H_s}) (\boldsymbol{\alpha}\circ\boldsymbol{y}) - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\ &+ \frac{1}{2\rho}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C})^{\mathrm{T}} (\boldsymbol{K^*} + \frac{2\lambda}{n\rho}\sum_{s=1}^n \boldsymbol{H_s^*}) (\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}) \\ s.t. \quad &\mathbf{1}^{\mathrm{T}}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}) = 0, \\ &\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0, \\ &0 \leq \alpha_i \leq \sigma' C, \\ &\beta_i \geq 0, \quad i = 1,\ldots,n.\end{aligned} \tag{14}$$

We further define two matrices $\boldsymbol{A} = \boldsymbol{K} + \frac{2\lambda}{n}\sum_{s=1}^n \boldsymbol{H_s}$ and $\boldsymbol{B} = \boldsymbol{K^*} + \frac{2\lambda}{n\rho}\sum_{s=1}^n \boldsymbol{H_s^*}$. And we let $\boldsymbol{\mu} = [\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2n}$, $\boldsymbol{v} = [(\mathbf{1} + \frac{1}{\rho}\boldsymbol{BC})^{\mathrm{T}}, (\frac{1}{\rho}\boldsymbol{BC})^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2n}$, and $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A}\circ(\boldsymbol{y}\boldsymbol{y}^{\mathbf{T}}) + \frac{1}{\rho}\boldsymbol{B} & \frac{1}{\rho}\boldsymbol{B} \\ \frac{1}{\rho}\boldsymbol{B} & \frac{1}{\rho}\boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$. Finally, the optimization problem of R-SVM+ can be rewritten as,

$$\begin{aligned}\min_{\boldsymbol{\mu}} \quad &\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\mu} \\ s.t. \quad &\mathbf{1}^{\mathrm{T}}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{C}) = 0, \\ &\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0, \\ &0 \leq \alpha_i \leq \sigma' C, \\ &\beta_i \geq 0, \quad i = 1,\ldots,n.\end{aligned} \tag{15}$$

Eq. (15) is a typical quadratic programming problem [Gould and Toint, 2004; Coleman and Li, 1996], which can be efficiently solved by off-the-shelf quadratic programming solvers.

4 Experiments

In order to evaluate the robustness of our proposed R-SVM+ algorithm, we carry out experiments on three real-world datasets for digit classification, face pose classification and human activity recognition tasks, respectively.

4.1 Experimental Setting

Datasets

The experiments are executed on three real-world datasets, including the MNIST+ dataset [Vapnik and Vashist, 2009], the RGB-D Face dataset [Hg *et al.*, 2012], and the Human Activity Recognition dataset [Anguita *et al.*, 2013]. The MNIST+

SNR	SVM	RSVM-RHHQ	SVM+	L2-SVM+	R-SVM+
4	82.09±1.49	80.26±4.02	81.76±2.71	81.07±2.67	**84.42±1.39**
6	83.17±1.78	83.37±1.58	82.96±3.26	82.76±1.96	**86.08±1.35**
8	84.64±1.63	84.59±1.79	84.74±2.78	84.12±1.84	**87.39±1.43**
10	85.81±1.55	85.86±1.62	86.03±2.80	85.47±1.55	**88.45±1.34**
15	87.15±1.76	86.10±1.41	88.82±2.06	88.06±1.69	**90.04±0.91**
No noise	90.97±0.94	87.83±0.84	91.86±0.70	92.12±0.70	**92.23±0.55**

Table 1: Classification accuracies (mean ± standard deviation, %) on the MNIST+ dataset. The best results on each row are in boldface.

	SVM	RSVM-RHHQ	SVM+	L2-SVM+	R-SVM+
Up vs Forward	95.07±1.22	93.33±1.95	95.63±1.69	95.87±1.82	**96.59±1.53**
Up vs Down	97.90±0.66	97.58±1.34	98.08±0.59	98.17±0.68	**98.48±0.56**
Up vs Expression	96.38±1.49	96.56±1.94	96.87±1.68	97.23±1.39	**97.72±1.34**
Forward vs Down	88.46±1.23	82.73±2.42	89.12±2.09	88.96±1.81	**91.63±2.01**
Forward vs Expression	88.73±2.81	85.60±2.26	88.33±1.80	88.88±1.90	**90.16±1.71**
Down vs Expression	89.19±2.41	87.72±2.14	89.55±2.11	90.04±2.18	**90.36±1.43**

Table 2: Classification accuracies (mean ± standard deviation, %) on the RGB-D Face dataset when SNR is equal to 8. The best results on each row are highlighted in boldface.

dataset is used for the digit classification task that classifies two digits "5" and "8". It contains 2943 images of "5" and 3025 images of "8" from the MNIST database [LeCun *et al.*, 1998]. A holistic (poetic) description [Vapnik and Vashist, 2009] for each image is translated into a 21-dimensional feature vector as privileged information. All the images of two digits in the MNIST+ dataset are resized into 10×10 pixels. The 100-dimensional vector of raw pixels is used as the example feature data for each image. The RGB-D Face dataset [Hg *et al.*, 2012] contains color and corresponding depth images of faces of 31 people in different face poses and expressions taken by a Kinect sensor. The depth images are used as privileged information. For each person, the image of each face pose is taken repeatedly for 3 times, which results in 31×3 RGB-depth image pairs for each pose. The face pose recognition task is performed on this dataset. The Human Activity Recognition dataset [Anguita *et al.*, 2013] contains 10299 instances of 30 people performing six activities (i.e., walking, walking upstairs, walking downstairs, sitting, standing, and laying) by wearing a smart phone on the waist. Each example for each activity is described by 561 dimensional features drawn from accelerometer, gyroscope, gravity signals and so on.

Implementation Details

For the MNIST+ dataset, it is randomly split into a training set of 100 images, a test set of 1866 images, and a validation set of 4002 [Vapnik and Vashist, 2009]. In the experiment, we randomly select 80 examples and their corresponding holistic descriptions from the training set as training examples for 10 times and classify the images on the test set.

For the RGB-D Face dataset, due to the small number of images per face pose, we merge the poses into four groups: looking up, looking forward, looking down and having facial expressions. Then we train a binary classifier on each pair of groups. We randomly split 40% color and corresponding depth image pairs per class for training, 30% image pairs per class for testing, and the rest 30% for validation for 10 times. We crop each image into the same fixed size of 150×150 and convert each color image into a gray image. Then for each image, it is divided into 100 non-overlapping subregions in 15×15 and for each subregion we extract the LBP feature. By concatenating the LBP features derived from all subregions, PCA is then performed to obtain a 150-dimensional compact representation.

For the Human Activity Recognition dataset, we use the first 200-dimensional features which come from the accelerometer and gyroscope 3-axial raw signals and their separation into body and gravity acceleration signals as example features. The remaining 361-dimensional features that come from signals obtained by some post-processing such as a Fast Fourier Transform (FFT) and the magnitude calculated using the Euclidean norm, are used as privileged features. We train one binary classifier on each pair of groups in the experiment. For training we use 200 examples from the desired class and 200 examples randomly drawn from the rest of examples from the remaining classes. And 600 examples randomly selected from the desired class and the rest of classes respectively are used for testing. The remaining examples from the desired class and the same number of examples from the rest of classes are used as the validation examples.

For each dataset, we add white Gaussian noise to examples in the validation set and test set with a specific signal-to-noise ratio (SNR). The classification results in experiments are averaged over 10 independent trials.

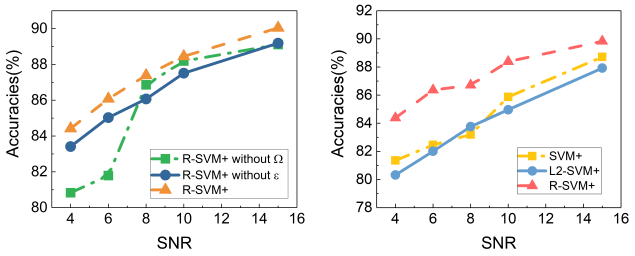
4.2 Compared Methods

For all the datasets, we evaluate performances of the proposed R-SVM+ algoirthm compared with the standard support vector machine (SVM), the robust SVM based on the rescaled hinge loss function (RSVM-RHHQ) [Xu *et al.*, 2016], SVM+ [Vapnik and Vashist, 2009], and L2-SVM+ [Li *et al.*, 2016] methods.

For all the methods, the regularization parameter C are selected from $10^{\{-2,1,0,1,2\}}$ and the Gaussian kernel is used.

	SVM	RSVM-RHHQ	SVM+	L2-SVM+	R-SVM+
Walking	89.52±2.17	92.07±1.21	91.90±1.16	92.21±1.37	**93.83±1.44**
Walking upstairs	82.38±1.90	91.88±0.98	92.36±0.49	89.98±2.23	**93.15±0.76**
Walking downstairs	92.04±1.30	93.71±0.68	94.29±0.86	94.37±0.52	**94.49±0.71**
Sitting	79.84±3.13	**87.75±0.87**	79.42±5.32	81.29±4.54	87.30±0.71
Standing	81.91±4.13	91.57±0.50	78.97±7.33	78.72±7.29	**91.61±1.00**
Laying	99.42±0.27	99.55±0.17	99.50±0.21	99.63±0.13	**99.65±0.12**

Table 3: Classification accuracies (mean ± standard deviation, %) on the Human Activity Recognition dataset with an SNR of 4. Best accuracies on each row are highlighted in boldface.



(a) Using separate components (b) Influence of noisy PF

Figure 1: Discussions of performances of R-SVM+.

For SVM+, L2-SVM+ and the proposed R-SVM+, we set the parameter of Gaussian kernel $\gamma = \frac{1}{D}$ where D is the mean of distances among examples in the training set according to [Li *et al.*, 2016]. While for SVM and RSVM-RHHQ, γ is selected from $10^{\{-3,-2,-1,0,1,2,3\}}$, since significantly better performances can be achieved. For RSVM-RHHQ, the scaling constant η is varied in range of $\{0.01, 0.1, 0.5, 1, 2, 3, 10, 100\}$. For SVM+-based methods, the trade-off parameter ρ is selected from $10^{\{-2,-1,0,1,2\}}$. For the proposed R-SVM+, we also vary the parameter σ' in range of $\{5, 10, 50, 100\}$ and λ in range of $10^{\{-5,-4,\ldots,0,1\}}$. The best parameters for all methods are determined with a joint cross validation model selection strategy on the validation set.

4.3 Performance Comparison

Table 1 summarizes the classification results of the proposed R-SVM+ algorithm and compared methods, using 80 clean training EF and PF examples, and noisy validation and test examples polluted by Gaussian noises with different SNRs of 4, 6, 8, 10 and 15. And we also report the classification results when validation and test examples are clean. As we can see, the performances of all the methods are significantly affected by varying SNR. Generally, the proposed R-SVM+ algorithm shows bigger superiority than all the compared methods especially when SNR is smaller. Since the distribution of noisy test data could be largely different from that of training data when SNR is small, the advantage of classical SVM+ over SVM is not preserved. When SNR is relatively large (i.e., SNR=15) or examples are clean, SVM+ and L2-SVM+ methods obviously outperform SVM and RSVM-RHHQ. While R-SVM+ achieves obviously better than SVM-based methods in all cases. This demonstrates the proposed R-SVM+ is a robust and effective SVM+-based algorithm against noises.

We also discuss the influences of different components of R-SVM+ on the classification results on the MNIST+ dataset, as an example. Figure 1 (a) shows the performances of R-SVM+ if we separately consider minimizing ϵ (referred as R-SVM+ without Ψ) and minimizing the value of $\Psi(t, \boldsymbol{\tau}_{\boldsymbol{x}}, \boldsymbol{\tau}_{\boldsymbol{x}^*})$ (referred as R-SVM+ without ϵ). We can find that in general the introduction of $\Psi(t, \boldsymbol{\tau}_{\boldsymbol{x}}, \boldsymbol{\tau}_{\boldsymbol{x}^*})$ obviously plays a more important role on the performance of R-SVM+ when SNR is small. While when SNR is relatively large, the introduction of ϵ will boost the accuracy to some extent. And R-SVM+ achieves the best performance by simultaneously considering both components in all cases.

In order to further evaluate the robustness of the proposed R-SVM+ algorithm compared with other SVM+-based methods when PF examples are noisy, we add Gaussian noises to the PF training examples with SNRs of 4, 6, 8, 10, and 15 on the MNIST+ dataset, as an example. As shown in Figure 1 (b), R-SVM+ obviously outperforms SVM+ and L2-SVM+ with varying SNR in general. In detail, R-SVM+ obtains gains in accuracy of +3.1%, +3.9%, +3.6%, +2.6% and +1.2% over SVM+, and gains in accuracy of +4.1%, +4.4%, +3.0%, +3.5% and +2.0% over L2-SVM+ when SNR is equal to 4, 6, 8, 10, and 15, respectively. This is because R-SVM+ also considers maximizing the lower bound of potential perturbations over PF examples and has more tolerances over these noisy examples, thus strengthening the robustness of the R-SVM+ algorithm.

Table 2 shows the performance of different methods on the RGB-D Face dataset when Gaussian noises with an SNR of 8 are added in the validation and test examples. Generally, we can observe that SVM+ and L2-SVM+ methods are superior to the SVM-based methods in almost all the 6 cases due to the use of depth images as privileged information. In particular, the proposed R-SVM+ algorithm shows its advantage over SVM+ and L2-SVM+ in all the 6 cases, especially in the forth and fifth cases.

Table 3 reports the classification results of the proposed R-SVM+ algorithm and compared methods on the Human Activity Recognition dataset whose validation and test examples are added with Gaussian noises with an SNR of 4. From the results shown in Table 3, we can see that R-SVM+ outperforms other methods in 5 out of 6 cases. In particular, R-SVM+ obtains obviously better accuracies than SVM, SVM+ and L2-SVM+ in the "Sitting" and "Standing" cases. This demonstrates the effectiveness and robustness of R-SVM+ against noises. And RSVM-RHHQ also performs obviously better than SVM because of its robust strategy based on SVM. In general, SVM+-based methods get better results than SVM because of the use of privileged information.

5 Conclusion

In this paper, we propose a Robust SVM+ (R-SVM+) algorithm to construct a more robust classifier in the LUPI paradigm for the potential noises in the data. Based on a rigorous theoretical analysis, the lower bound of perturbations of noises that will mislead the model to make incorrect decisions has been evaluated. Accordingly, R-SVM+ can be learned by introducing a novel regularization function into a slack form of SVM+. The effectiveness of the proposed R-SVM+ algorithm and the necessity of studying robust SVM+ methods in the LUPI paradigm are demonstrated by experiments on real-world datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China under grants U1536204, 61471274, and 61711530239, and Australian Research Council Projects: FL-170100117, DE-180101438, DP-180103424, and LP-150100671.

References

[Anguita *et al.*, 2013] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J L Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.

[Coleman and Li, 1996] Thomas F. Coleman and Yuying Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058, 1996.

[Feyereisl *et al.*, 2014] Jan Feyereisl, Suha Kwak, Jeany Son, and Bohyung Han. Object localization based on structural SVM using privileged information. In *Advances in Neural Information Processing Systems*, pages 208–216, 2014.

[Gould and Toint, 2004] Nick Gould and Philippe L. Toint. *Preprocessing for quadratic programming*. Springer-Verlag New York, Inc., 2004.

[Hg *et al.*, 2012] R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet. An rgb-d database using microsoft's kinect for windows for face detection. In *Eighth International Conference on Signal Image Technology and Internet Based Systems*, pages 42–46, 2012.

[Ji *et al.*, 2012] You Ji, Shiliang Sun, and Yue Lu. Multitask multiclass privileged information support vector machines. In *2012 21st International Conference on Pattern Recognition*, pages 2323–2326. IEEE, 2012.

[LeCun *et al.*, 1998] Y. L. LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. proc ieee. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2016] Wen Li, Dengxin Dai, Mingkui Tan, Dong Xu, and Luc Van Gool. Fast algorithms for linear and kernel SVM+. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2258–2266, 2016.

[Motiian *et al.*, 2016] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Information bottleneck learning using privileged information for visual recognition. In *Computer Vision and Pattern Recognition*, pages 1496–1505, 2016.

[Niu *et al.*, 2012] Lingfeng Niu, Yong Shi, and Jianmin Wu. Learning using privileged information with L-1 support vector machine. In *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 10–14, 2012.

[Pechyony and Vapnik, 2010] Dmitry Pechyony and Vladimir Vapnik. On the theory of learning with privileged information. In *International Conference on Neural Information Processing Systems*, pages 1894–1902, 2010.

[Sharmanska *et al.*, 2013] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 825–832, 2013.

[Shen *et al.*, 2014] Jianbing Shen, Yunfan Du, Wenguan Wang, and Xuelong Li. Lazy random walks for superpixel segmentation. *IEEE Transactions on Image Processing*, 23(4):1451–1462, 2014.

[Vapnik and Izmailov, 2015] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(55):2023–2049, 2015.

[Vapnik and Vashist, 2009] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009.

[Wang *et al.*, 2018] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2018.

[Xu *et al.*, 2016] Guibiao Xu, Zheng Cao, Bao Gang Hu, and Jose C. Principe. Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognition*, 63:139–148, 2016.

[Yang *et al.*, 2017] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017.

[You *et al.*, 2017] Shan You, Chang Xu, Yunhe Wang, Chao Xu, and Dacheng Tao. Privileged multi-label learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3336–3342, 2017.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Xinxing Xu, Sinno Jialin Pan, Ivor W Tsang, Zheng Qin, and Rick Siow Mong Goh. Transfer hashing with privileged information. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2414–2420, 2016.