Robust Learning with Imperfect Privileged Information

Xue Li\textsuperscript{a,b}, Bo Du\textsuperscript{a,∗}, Chang Xu\textsuperscript{c}, Yipeng Zhang\textsuperscript{a}, Lefei Zhang\textsuperscript{a}, Dacheng Tao\textsuperscript{c}

\textsuperscript{a}School of Computer Science, Wuhan University, China
\textsuperscript{b}State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China
\textsuperscript{c}UBTECH Sydney Artificial Intelligence Centre and the School of Computer Science, the Faculty of Engineering and Information Technologies, the University of Sydney, Australia

Abstract

In the learning using privileged information (LUPI) paradigm, example data cannot always be clean and gathered privileged information could be imperfect in practice. Imperfect privileged information can be privileged information which is not always accurate or perturbed by noise, or incomplete privileged information that privileged information is only available for part of training data. Because of the lack of clear strategies to handle noise in the example data and imperfect privileged information, existing learning using privileged information (LUPI) methods may be put into a dilemma and their performance could be seriously challenged. In this paper, we propose a Robust SVM+ method to tackle imperfect data in LUPI. In order to make the SVM+ model robust to the noises of example data and privileged information, Robust SVM+ maximizes the lower bound of the perturbations that may influence the judgement based on a rigorous theoretical analysis. Moreover, to deal with the incomplete privileged information, we employ the available privileged information for help to approximate the missing privileged information of training data. The optimiza-

\textsuperscript{∗}This paper extends the work presented at the 27th International Joint Conference on Artificial Intelligence, IJCAI-18 (paper titled “R-SVM+: Robust Learning with Privileged Information”).

\textsuperscript{∗}Corresponding author

Email addresses: lixue93@whu.edu.cn (Xue Li), remoteking@whu.edu.cn (Bo Du), c.xu@sydney.edu.au (Chang Xu), zyp91@whu.edu.cn (Yipeng Zhang), zhanglefei@whu.edu.cn (Lefei Zhang), dacheng.tao@sydney.edu.au (Dacheng Tao)
tion problem of the proposed method can be efficiently solved by a two-step alternating optimization strategy based on off-the-shelf quadratic programming solvers and the alternating direction method of multipliers (ADMM) technique iteratively. Comprehensive experiments on real-world datasets demonstrate the effectiveness of the proposed Robust SVM+ method to handle imperfect privileged information.

**Keywords:** Learning using privileged information, Classification, Support vector machine

1. Introduction

Recently, the advances of machine learning have brought about great influence on the development of artificial intelligence [1, 2, 3]. In real-world applications, supervised learning [4] is one of major machine learning paradigms [5, 6] to explore and exploit information underlying the data to understand the world. Traditional supervised learning methods aim to classify unknown test examples using a classification model, which has been well trained, with a set of observed examples and their corresponding labels. These supervised learning methods act like students who learn knowledge and skills from textbooks with questions and answers. However, in human teaching and learning process, it is a widespread consensus that the role of a teacher is very important for the learning of students. This depicts a fact that a student can learn faster and better with the help of a teacher which can supply some useful information such as comments, explanations, comparisons and so on. If a teacher can be also considered in machine learning process, it may make a difference.

Based on this insightful observation, Vapnik and Vashist [7] introduced the learning using privileged information (LUPI) paradigm, and made a success by using the auxiliary information supplied by a teacher to help improve the model learning. It is noted that the auxiliary information is only available at the training stage and not at the test stage, thus it is referred to as privileged information. Because a teacher can provide information and knowledge for students
in the learning process, while during the examination or test a teacher will not help at all. Privileged information is common and useful in practice [8, 9, 10]. For example, in the task of action recognition from depth sequences, skeleton joints can be considered as privileged information to provide more knowledge to improve generalizability and avoid overfitting the scarcer training data [11]. When classifying the images of animal categories, the attribute annotations for the animals can be captured as privileged information to mitigate the lack of good quality labeled data [12].

Recently, LUPI has attracted more and more attentions in different applications and been investigated in many existing algorithms [13, 14, 15, 16, 17]. One of the most classical LUPI method is SVM+ [7]. It defines a correcting function to learn the slack variables with the help of privileged information in the framework of the support vector machine (SVM) technique [18]. Due to the success of SVM+, there are several works have been studied based on SVM classifiers [19, 20, 21, 22, 23]. Despite these works based on SVM, privileged information has also been exploited in different algorithm frameworks [24, 25]. In convolutional neural networks (CNNs) and recurrent neural networks (RNNs), a heteroscedastic dropout was developed by taking the variance of the dropout as a function of privileged information [26]. A privileged matrix factorization method is proposed to utilize review texts to assist the learning of user and item factors for collaborative filtering [27]. A modality hallucination architecture for RGB object detection is introduced by incorporating depth privileged information [28].

Most of these works are developed in an ideal scenario, where all training examples are supplied with the corresponding privileged information, example data are deemed to be clean and the privileged information provided by a teacher is always perfect. However, in the real applications, this ideal case cannot always be met. We can encourage the model to be well trained on the observed training set. But at the test stage, it is hard and even impossible to tell what test data will be. Moreover, it is also difficult to determine whether the available privileged information is always accurate, in good-quality, and each
example training data is supplied with privileged information. If not, these
imperfect privileged information will influence the teacher’s correct comments
on the learning of LUPI model. Though there are a few works investigate the
incomplete privileged information, they tend to only use the part of training
elements with privileged information [7] [29]. That is, examples with privileged
information are exploited in the LUPI way, while examples without privileged
information are used in the conventional supervised way. They do not con-
sider how to fully exploit the part of available privileged information to learn
the incomplete privileged information and has no consideration about the more
general imperfect privileged information issues.

In this paper, we propose a Robust SVM+ method for noisy example data
and imperfect privileged information. It can construct a more robust classifier
based on the SVM+ model, by analyzing the lower bound of perturbations of ex-
ample data and privileged information that will mislead the decision of SVM+.
Moreover, in order to fully use the available privileged information to learn the
incomplete privileged information, based on a novel Incomplete SVM+ model,
Robust SVM+ is also extended to incomplete privileged information situation.

To the best of our knowledge, it is the first attempt to try to re-construct incom-
plete privileged information for examples using available privileged information.
Specifically, we define a dictionary where atoms are the feature vectors in the
reproducing kernel Hilbert space (RKHS) induced by the kernel on available
privileged feature data. Since each data point in a union of subspaces can be ef-
ficiently represented as a linear combination of other points [30], it is reasonable
to re-construct the incomplete privileged feature data by a linear combination
of the atoms in the dictionary. This paper is an extension of our preliminary
work [31]. In this paper, we extend the preliminary work into the more gener-
ally imperfect privileged information situation including incomplete privileged
information, based on a novel Incomplete SVM+ framework. Moreover, the
proposed Robust SVM+ method can deal with both noisy example data and
imperfect privileged information. Experimental results on real-world datasets
demonstrate the effectiveness of the proposed Robust SVM+ method.
The rest of this paper is organized as follows. Section 2 presents necessary background and preliminary of SVM+. Section 3 and 4 describe the proposed methods for noisy example data and imperfect privileged information including privileged information is not always accurate or perturbed by noise, and incomplete privileged information where privileged information is only available for part of training data. Section 5 presents the optimization of the proposed method. Section 6 evaluates the propose methods compared with several competing methods on several real-world datasets. Section 7 concludes the paper.

2. Preliminary

Let \( x_i \in \mathbb{R}^d \) denote the \( i \)-th example feature (EF) vector and \( y_i \in \{+1, -1\} \) denote the corresponding ground-truth label. \( x_i^* \in \mathbb{R}^{d^*} \) is the privileged feature (PF) vector corresponding to \( x_i \). The conventional supervised learning exploits a set of \( n \) example feature vectors \( \{x_1, \ldots, x_n\} \) and the corresponding labels \( \{y_1, \ldots, y_n\} \) at the training stage. While the LUPI paradigm requires a set of \( n \) training example pairs where privileged information is additionally provided, \( (x_1^*, y_1), \ldots, (x_i^*, y_i), \ldots, (x_n^*, y_n) \).

A typical LUPI method is SVM+ which estimates the slack variables in SVM by defining a correcting function with privileged information. Its objective function is formulated as follows:

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

(1)

where \( w \) and \( b \) respectively denote the weight vector and the bias term in the example feature space, and \( w^* \) and \( b^* \) are respectively the weight vector and the bias term in the privileged feature space. \( \phi(\cdot) \) and \( \psi(\cdot) \) denote two feature mapping functions which are induced by the kernels on example features and
privileged features, respectively. $\rho > 0$ and $C > 0$ are the trade-off parameters. $\langle a, e \rangle$ is the inner product of two vectors $a$ and $e$.

3. Robust SVM+

In this section, we consider the situation where example data are not clean with noise and the collected privileged information data are not always accurate or perturbed by noise. Based on the constraint of Eq. (1), we can define two functions $f(x) = 1 - y[(w, \phi(x)) + b]$ and $g(x^*) = (w^*, \psi(x^*)) + b^*$. Then the constraint $y_i[(w, \phi(x_i)) + b] \geq 1 - [(w^*, \psi(x_i^*)) + b^*]$ in Eq. (1) can be transformed to

$$f(x) \leq g(x^*),$$

which will be satisfied if $x$ and $x^*$ are good enough for model training and a small loss in the decision space will be achieved. It is also a basic assumption of SVM+ that a small loss in the decision space should be obtained, if a small loss in the correcting space can be achieved [32].

When the perturbations $\tau_x \in \mathbb{R}^d$ caused by noises over ideal observations $x$ and the perturbations $\tau_{x^*} \in \mathbb{R}^{d^*}$ of collected privileged information over ideal privileged information $x^*$ provided by a teacher are large enough, the inequality will not hold and thus will influence the correct decisions of the model. The inequality will become

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

In the following, we derive the objective function of Robust SVM+ with complete privileged information (referred as R-SVM+) by analyzing the lower bounds of $\tau_x$ and $\tau_{x^*}$ through a rigorous theoretical analysis. Based on the fundamental theorem of calculus, we can obtain

$$f(x + \tau_x) = f(x) + \int_0^1 \langle \nabla f(x + t\tau_x), \tau_x \rangle dt,$$

$$g(x^* + \tau_{x^*}) = g(x^*) + \int_0^1 \langle \nabla g(x^* + t\tau_{x^*}), \tau_{x^*} \rangle dt.$$
If the perturbations are serious enough, according to Eq. (3) and Eq. (4), we can obtain the following inequality,

\[
0 \leq g(x^*) - f(x) < \int_0^1 \langle \nabla f(x + t\tau_x), \tau_x \rangle dt - \int_0^1 \langle \nabla g(x^* + t\tau_{x^*}), \tau_{x^*} \rangle dt = \int_0^1 \left[ \nabla f(x + t\tau_x); -\nabla g(x^* + t\tau_{x^*}) \right]^T [\tau_x; \tau_{x^*}] dt \leq \|\tau\|_p \int_0^1 \|\varpi(t, \tau_x, \tau_{x^*})\|_q dt, \tag{5}
\]

where \(\tau = [\tau_x; \tau_{x^*}] \in \mathbb{R}^{d + d^*}\), and \(\varpi(t, \tau_x, \tau_{x^*}) = [\nabla f(x + t\tau_x); -\nabla g(x^* + t\tau_{x^*})] \in \mathbb{R}^{d + d^*}\). The final inequality can be achieved using Hölder’s inequality, where \(p \in [1, +\infty]\) and \(q \in [1, +\infty]\) satisfy \(\frac{1}{p} + \frac{1}{q} = 1\). In the following, we discuss the case \(p = q = 2\). Given \(g(x^*) \geq f(x)\), we can obtain the minimal perturbations \(\tau\) that will reverse the decision of the SVM+ classifier.

\[
\|\tau\|_2 > \frac{g(x^*) - f(x)}{\int_0^1 \|\varpi(t, \tau_x, \tau_{x^*})\|_2 dt}. \tag{6}
\]

According to Eq. (6), the lower bound (i.e., the right of Eq. (6)) should be maximized to obtain a more robust classifier. Therefore, the R-SVM+ algorithm considers making \(\int_0^1 \|\varpi(t, \tau_x, \tau_{x^*})\|_2 dt\) small and \(g(x^*) - f(x)\) large. Thus, the objective function of R-SVM+ have two components: 1) for \(f(x) \leq g(x^*) + \epsilon\), \(\epsilon\) should be minimized; 2) minimizing the value of function \(\Psi(t, \tau_x, \tau_{x^*}) = \int_0^1 \|\varpi(t, \tau_x, \tau_{x^*})\|_2 dt\).

3.1. Minimizing \(\epsilon\)

In order to minimize \(\epsilon\), we replace the constraints \(f(x_i) \leq g(x_i^*)\), \(i = 1, \ldots, n\) in SVM+ with the constraints \(f(x_i) \leq g(x_i^*) + \epsilon_i\), where \(\epsilon_i \geq 0, i = 1, \ldots, n\). And the objective function of this variant form of the SVM+ problem can be
transformed into,
\[
\begin{align*}
\min_{w, w^*, b, b^*, \epsilon} & \quad \frac{1}{2} \langle w, w \rangle + \rho \langle w^*, w^* \rangle \\
& + C \sum_{i=1}^{n} [\langle w^*, \psi(x_i^*) \rangle + b^*] + \sigma \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} & \quad y_i [(w, \phi(x_i)) + b] \geq 1 - [\langle w^*, \psi(x_i^*) \rangle + b^*] - \epsilon_i, \\
& \quad \langle w^*, \psi(x_i^*) \rangle + b^* \geq 0, \\
& \quad \epsilon_i \geq 0, \quad i = 1, \ldots, n,
\end{align*}
\]

where \( \sigma > 0 \) is a tradeoff parameter. Similar variant of Eq. (7) has been discussed in [7] as well, while here the value of \( g(x^*) - f(x) \) is made to be large to enhance the robustness according to Eq. (6). By introducing Lagrange multipliers \( \alpha_i \geq 0, \beta_i \geq 0 \) and \( \eta_i \geq 0 \), where \( i = 1, \ldots, n \), the dual problem of Eq. (7) can be written as,
\[
\begin{align*}
\max_{\alpha, \beta} & \quad 1^T \alpha - \frac{1}{2} (\alpha \circ y)^T K (\alpha \circ y) \\
& - \frac{1}{2 \rho} (\alpha + \beta - C)^T K^* (\alpha + \beta - C), \\
\text{s.t.} & \quad 1^T (\alpha + \beta - C) = 0, \\
& \quad y^T \alpha = 0, \\
& \quad 0 \leq \alpha \leq \sigma, \\
& \quad \beta \geq 0.
\end{align*}
\]

For expression simplicity, we define \( \sigma = \sigma' C \) where \( \sigma' > 1 \). \( \beta = [\beta_1, \ldots, \beta_n]^T \in \mathbb{R}^n \), \( 1 = [1, \ldots, 1]^T \in \mathbb{R}^n \), and \( C = [C, \ldots, C]^T \in \mathbb{R}^n \). \( \alpha \circ y \) is denoted as the element-wise product between vectors \( \alpha = [\alpha_1, \ldots, \alpha_n]^T \in \mathbb{R}^n \) and \( y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n \). Moreover, \( K \in \mathbb{R}^{n \times n} \) is a kernel matrix where \( K_{ij} = k(x_i, x_j) \) and \( K^* \) is the kernel matrix of privileged features where \( K^*_{ij} = k^*(x_i^*, x_j^*) \in \mathbb{R}^{n \times n} \).

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

We assume that \( \tau_x \) and \( \tau_{x^*} \) are within some fixed range that satisfies \( \tau_x, \tau_{x^*} \in \Omega_p(0, \ell) \), where \( \Omega_p(x, \ell) = \{ z \in \mathbb{R}^d \| x - z \|_2 \leq \ell \} \). Then an ex-
ample with noisy features and its inaccurate privileged information become $z = x + t\tau_x$ and $z^* = x^* + t\tau^*$, $0 < t \leq 1$. To minimize $\Psi(t, \tau_x, \tau^*) = \int_0^1 \|\nabla(t, \tau_x, \tau^*)\|_2 dt$, we can obtain

$$\sup_{\tau_x, \tau^*} \int_0^1 \|\nabla(t, \tau_x, \tau^*)\|_2 dt \leq \max_{x \in \Omega_2(0, \ell)} \|\nabla f(x); -\nabla g(x^*)\|_2. \tag{9}$$

Therefore, the problem can be transformed into the minimization problem with respect to the value of the upper bound, i.e., the right side of Eq. (9). To simplify the optimization problem, the mean of $\|\nabla f(x); -\nabla g(x^*)\|_2^2$ on all the training examples with privileged information are used to substitute the upper bound. Thus we define

$$\Theta(f, g) = \frac{1}{n} \sum_{s=1}^n \|\nabla f(x_s); -\nabla g(x^*_s)\|_2^2$$

$$= \frac{1}{n} \sum_{s=1}^n \| - y_s \sum_{i=1}^n \alpha_i y_i \nabla x_i k(x_i, x_s)\|_2^2$$

$$+ \frac{1}{n} \sum_{s=1}^n \| - \frac{1}{\rho} \sum_{i=1}^n (\alpha_i + \beta_i - C) \nabla x_i k(x_i^*, x_s)\|_2^2$$

$$= \frac{1}{n} \sum_{s=1}^n (\alpha \circ y)^T H_s (\alpha \circ y) + \frac{1}{n\rho^2} \sum_{s=1}^n (\alpha + \beta - C)^T H_s^* (\alpha + \beta - C), \tag{10}$$

where

$$h(x_i, x_j, x_s) = \langle \nabla x_i k(x_i, x_s), \nabla x_j k(x_j, x_s) \rangle = 4\gamma^2 \langle x_i - x_s, x_i - x_j \rangle e^{-\gamma \|x_i - x_s\|^2} - \gamma \|x_i - x_s\|^2,$$

and

$$h^*(x_i^*, x_j^*, x_s^*) = \langle \nabla x_i^* k(x_i^*, x_s), \nabla x_j^* k(x_j^*, x_s) \rangle = 4\gamma^2 \langle x_i^* - x_s^*, x_i^* - x_j^* \rangle e^{-\gamma \|x_i^* - x_s^*\|^2} - \gamma \|x_i^* - x_s^*\|^2,$$

and $H_s$ is defined as a matrix with each element being $H_{s,ij} = h(x_i, x_j, x_s) \in \mathbb{R}^{n \times n}$ and $H_s^*$ is defined as a matrix with each element being $H_{s,ij}^* = h^*(x_i, x_j, x_s) \in \mathbb{R}^{n \times n}$. Similar variant of Eq. (10) has been discussed in [33] which tries to make the differences of classifier functions of example data points constant in the traditional supervised paradigm. Differently, Eq. (10) aims to restrict functions in
both example feature and privileged feature spaces under the LUPI paradigm, to learn a robust model with more tolerances over noisy data and inaccurate privileged information. And the problem can be finally written as,
\[
\min_{\alpha, \beta} \Theta(f, g).
\] (11)

3.3. Objective Function

We derive the objective of R-SVM+ which considers a maximization problem in Eq. (8) and a minimization problem in Eq. (11) at the same time,
\[
\min_{\alpha, \beta} \frac{1}{2} (\alpha \circ y)^T K (\alpha \circ y) - 1^T \alpha \\
+ \frac{1}{2\rho} (\alpha + \beta - C)^T K^* (\alpha + \beta - C) + \lambda \Theta(f, g),
\] s.t. \[ 1^T (\alpha + \beta - C) = 0, \]
\[ y^T \alpha = 0, \]
\[ 0 \leq \alpha_i \leq \sigma'C, \]
\[ \beta_i \geq 0, \quad i = 1, \ldots, n. \] (12)

And \( \lambda > 0 \) is a trade-off parameter. By defining two matrices \( A = K + \frac{2\lambda}{n} \sum_{s=1}^{n} H_s \) and \( B = K^* + \frac{2\lambda}{n\rho} \sum_{s=1}^{n} H^*_s \), \( \mu = [\alpha^T, \beta^T]^T \in \mathbb{R}^{2n} \), \( v = [(1 + \frac{1}{\rho} BC)^T, (\frac{1}{\rho} BC)^T]^T \in \mathbb{R}^{2n} \), and \( M = \begin{bmatrix} A \circ (yy^T) + \frac{1}{\rho} B & \frac{1}{\rho} B \\
\frac{1}{\rho} B & \frac{1}{\rho} B \end{bmatrix} \in \mathbb{R}^{2n \times 2n} \), the objective can be rewritten as
\[
\min_{\mu} \frac{1}{2} \mu^T M \mu - v^T \mu
\] s.t. \[ 1^T (\alpha + \beta - C) = 0, \]
\[ y^T \alpha = 0, \]
\[ 0 \leq \alpha_i \leq \sigma'C, \]
\[ \beta_i \geq 0, \quad i = 1, \ldots, n. \] (13)

Eq. (13) is a typical quadratic programming problem \[34\] \[35\], and it can be efficiently solved by off-the-shelf quadratic programming solvers.
4. Incomplete Privileged Information

In the section, we consider the situation where privileged information is available only for a part of examples and extend the proposed Robust SVM+ to deal with this situation. Let us consider a set of training examples,

\((x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell), (x_{\ell+1}, y_{\ell+1}), \ldots, (x_n, y_n)\),

where only \(\ell\) examples are supplied with privileged information, while the remaining \((n - \ell)\) examples have no privileged information.

According to the observations that each data point in a union of subspaces can be efficiently represented as a linear combination of other points \([30]\), it is reasonable to consider learning and re-constructing the incomplete privileged information \(\{x_\ell^*+1, \ldots, x_n^*\}\) based on the linear combination of existing privileged information. That is, we construct a dictionary \(D^* = [\psi(x_1^*), \ldots, \psi(x_\ell^*)]\), where the functions \(\psi(\cdot)\) is a feature mapping function in the reproducing kernel Hilbert space (RKHS) induced by the kernel on privileged feature. Then we can estimate and re-construct the feature representations of incomplete privileged information in RKHS with \(\psi(x_j^*) = D^*v_j\), where \(v_j\) is a coefficient vector and \(j = \ell + 1, \ldots, n\). And the feature vectors of the available privileged data can also be represented as \(\psi(x_i^*) = D^*v_i\) where \(i = 1, \ldots, \ell\) and \(v_i\) is a fixed vector whose the \(i\)-th element is 1 while other elements are 0. We first propose an Incomplete SVM+ model, whose objective function with respect to variables \(w, w^*, b, b^*\) and \(v_j, j = \ell + 1, \ldots, n\) becomes:

\[
\min_{v_j \geq 0, j=\ell+1,\ldots,n} \min_{w, w^*, b, b^*} \frac{1}{2} \langle w, w \rangle + \rho \langle w^*, w^* \rangle + C_1 \sum_{i=1}^{n} [\langle w^*, D^*v_i \rangle + b^*] + C_2 \sum_{j=\ell+1}^{n} \|1_\ell v_j - 1\|_2^2
\]

\[\quad \text{s.t.} \quad y_i [\langle w, \phi(x_i) \rangle + b] \geq 1 - [\langle w^*, D^*v_i \rangle + b^*], \]

\[\quad \langle w^*, D^*v_i \rangle + b^* \geq 0, \quad i = 1, \ldots, n.\]

where \(\sum_{j=\ell+1}^{n} \|1_\ell v_j - 1\|_2^2\) is a sum-to-one regularization term, \(C_1\) and \(C_2\) are trade-off parameters, and \(1_\ell = [1, \ldots, 1]^T \in \mathbb{R}^\ell\).
We define two functions $f(x)$ and $\zeta(v)$. $f(x) = 1 - y[\langle w, \phi(x) \rangle + b]$ and $\zeta(v) = \langle w^*, D^*v \rangle + b^*$. Then the first constraint of Eq. (14) can be rewritten as $f(x) \leq \zeta(v)$. Considering the situation in practice when noises exist in data and the estimated $v_i$ may not always be so accurate, we need to learn a more robust model which can reduce the influence of some inaccurate learned $v_i$ on the decision. We define observations of example feature data in reality as $x + \tau_x$ and evaluated coefficient vector as $v + \tau_v$. $\tau_x \in \mathbb{R}^d$ and $\tau_v \in \mathbb{R}^l$ denote the perturbations over $x$ and $v$, respectively. Similar to the calculation process of Eqs. (2)-(5), we can finally obtain the minimal perturbations $\tau = [\tau_x; \tau_v]$ which will disturb the correct decision of Eq. (14) classifier as follows,

$$\|\tau\|_2 > \frac{\zeta(v) - f(x)}{\Gamma(\tau_x, \tau_v)}.$$  

(15)

where $\Gamma(\tau_x, \tau_v) = \int_0^1 \|\delta(t, \tau_x, \tau_v)\|_2 dt$ and $\delta(t, \tau_x, \tau_v) = [\nabla f(x + t\tau_x); -\nabla \zeta(v + t\tau_v)]$. Based on Eq. (15), we consider maximizing the right of the inequality to enhance the robustness of the Incomplete SVM+ classifier. In this way, the derived model can tolerate relatively more perturbations in example data and re-constructed privileged data without reversing the decision. Therefore, Robust SVM+ considers making 1) $\zeta(v) - f(x)$ large and 2) $\Gamma(\tau_x, \tau_v)$ small at the same time. For the first subproblem, we consider minimizing an intermediate variable $\varepsilon$ which satisfies $f(x) \leq \zeta(v) + \varepsilon$.

4.1. Minimizing $\varepsilon$

In order to minimize the intermediate variable $\varepsilon$, we replace the original first constraint of the objective function in Eq. (14) with the inequality $f(x) \leq \zeta(v) + \varepsilon$, and add a minimization term related to $\varepsilon$ in the objective. The variant of Robust SVM+ can be derived,
\[
\begin{align*}
\min_{v_j \geq 0, j = \ell + 1, \ldots, n} & \quad \min_{w, w^*, b, b^*, \varepsilon_i} \\
& \quad \frac{1}{2}(\langle w, w \rangle + \rho(\langle w^*, w^* \rangle)) + C_1 \sum_{i=1}^{n} [\langle w^*, D^*v_i \rangle + b^*] \\
& \quad + C_2 \sum_{j=\ell+1}^{n} \|1\ell v_j - 1\|^2_2 + C_3 \sum_{i=1}^{n} \varepsilon_i, \\
\text{s.t.} & \quad y_i(\langle w, \phi(x_i) \rangle + b) \geq 1 - [\langle w^*, D^*v_i \rangle + b^*] - \varepsilon_i, \\
& \quad \langle w^*, D^*v_i \rangle + b^* \geq 0, \\
& \quad \varepsilon_i \geq 0, \quad i = 1, \ldots, n.
\end{align*}
\]

where \(C_3\) is a non-negative trade-off parameter. For simplicity, we let \(C_3 = \sigma' C_1\) where \(\sigma' > 1\). Since it is difficult to solve Eq. (16) directly, we first derive the dual problem of its inner minimization problem related to \(w, w^*, b, b^*, \varepsilon_i\). In detail, we introduce Lagrange multipliers \(\alpha_i \geq 0, \beta_i \geq 0, \) and \(\eta_i \geq 0\) where \(i = 1, \ldots, n\) for the constraints in Eq. (16). Then, the Lagrangian function can be constructed as

\[
L = \frac{1}{2}(\langle w, w \rangle + \rho(\langle w^*, w^* \rangle)) + C_1 \sum_{i=1}^{n} [\langle w^*, D^*v_i \rangle + b^*] + C_2 \sum_{j=\ell+1}^{n} \|1\ell v_j - 1\|^2_2 + C_3 \sum_{i=1}^{n} \varepsilon_i - \sum_{i=1}^{n} \alpha_i \{y_i[\langle w, \phi(x_i) \rangle + b] - 1 + [\langle w^*, D^*v_i \rangle] + b^*] - \varepsilon_i\}
- \sum_{i=1}^{n} \beta_i[\langle w^*, \psi(x_i^*) \rangle + b^*] - \sum_{i=1}^{n} \eta_i \varepsilon_i.
\]

By setting derivatives of \(L\) with respect to \(w, w^*, b, b^*\) and \(\varepsilon_i\) to zero, we can obtain Karush-Kuhn-Tucker (KKT) conditions

\[
\begin{align*}
w &= \sum_{i=1}^{n} \alpha_i y_i \phi(x_i), \\
w^* &= \frac{1}{\rho} \sum_{i=1}^{n} (\alpha_i + \beta_i - C_1) D^*v_i, \\
\sum_{i=1}^{n} \alpha_i y_i &= 0,
\end{align*}
\]
\[
\sum_{i=1}^{n} (\alpha_i + \beta_i - C_1) = 0, \\
0 \leq \alpha_i \leq C_3, \quad \beta_i \geq 0, \quad i = 1, \ldots, n.
\]

With the KKT conditions, the dual problem can be formulated as:

\[
\begin{align*}
\min_{v_j \geq 0} \quad & \max_{\alpha, \beta} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + C_2 \sum_{j=\ell+1}^{n} \|1_{\ell}v_j - 1\|_2^2 \\
& \quad - \frac{1}{2p} \sum_{i,j=1}^{n} (\alpha_i + \beta_i - C_1)(\alpha_j + \beta_j - C_1) v_i^T D^* D^* v_j, \\
\text{s.t.} & \quad \sum_{i=1}^{n} (\alpha_i + \beta_i - C_1) = 0, \\
& \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \\
& \quad 0 \leq \alpha_i \leq C_3, \\
& \quad \beta_i \geq 0, \quad i = 1, \ldots, n,
\end{align*}
\]

We further define vectors \(\alpha = [\alpha_1, \ldots, \alpha_n]^T \in \mathbb{R}^n\), \(y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n\), \(\beta = [\beta_1, \ldots, \beta_n]^T \in \mathbb{R}^n\), \(1_n = [1, \ldots, 1]^T \in \mathbb{R}^n\), \(C_1 = [C_1, \ldots, C_1]^T \in \mathbb{R}^n\) and \(C_3 = [C_3, \ldots, C_3]^T \in \mathbb{R}^n\). And we also define \(V = [v_1, \ldots, v_n] \in \mathbb{R}^{\ell \times n}\) and \(\tilde{V} = [v_{\ell+1}, \ldots, v_n] \in \mathbb{R}^{(n-\ell) \times n}\). The dual formulation can be further reformulated as:

\[
\begin{align*}
\min_{v_j \geq 0} \quad & \max_{\alpha, \beta} \quad 1^T \alpha - \frac{1}{2} (\alpha \circ y)^T K (\alpha \circ y) + C_2 \|1_{\ell}^T \tilde{V} - 1_n^T\|_2^2 \\
& \quad - \frac{1}{2p} (\alpha + \beta - C_1)^T V^T D^* D^* V (\alpha + \beta - C_1), \\
\text{s.t.} & \quad 1^T (\alpha + \beta - C_1) = 0, \\
& \quad y^T \alpha = 0, \\
& \quad 0 \leq \alpha \leq C_3, \\
& \quad \beta \geq 0.
\end{align*}
\]

where \(K \in \mathbb{R}^{n \times n}\) denotes the kernel matrix whose each element is \(K_{ij} = k(x_i, x_j)\) and \(\alpha \circ y\) represents the element-wise product between vectors \(\alpha\) and \(y\). 

(18)
4.2. Minimizing $\Gamma(\tau_x, \tau_v)$

We assume that perturbations caused by noises and some inaccurate estimations of coefficient vector $v$ are within certain ranges, i.e., $\tau_x \in \Omega_2(0, r)$ and $\tau_v \in \Omega_2(0, r)$, where $\Omega_2(a, r) = \{z \in \mathbb{R}^d ||a - z||_2 \leq r, r > 0\}$. Then we define $z = x + t\tau_x$ and $v^* = v + t\tau_v$. It is easy to obtain Eq. (20):

$$
\sup_{\tau_x, \tau_v \in \Omega_2(0, r)} \int_0^1 ||\delta(t, \tau_x, \tau_v)||_2 dt \leq \max_{\tau_x, \tau_v \in \Omega_2(0, r)} ||\delta(t, \tau_x, \tau_v)||_2 \\
= \max_{z \in \Omega_2(x, r), v^* \in \Omega_2(v, r)} ||\nabla f(z); -\nabla \zeta(v^*)||_2.
$$

The problem of minimizing $\Gamma(\tau_x, \tau_v)$ can be transformed into minimizing the upper bound of $\Gamma(\tau_x, \tau_v)$ in Eq. (20). In order to simplify the problem, we use the mean of $||\nabla f(z); -\nabla \zeta(v^*)||_2^2$ on all the example data points in the training set to substitute the upper bound. In detail, we define a function

$$
\Theta(f, \zeta) = \frac{1}{n} \sum_{s=1}^{n} ||\nabla f(x_s); -\nabla \zeta(v_s)||_2^2 \\
= \frac{1}{n} \sum_{s=1}^{n} \left| -y_s \sum_{i=1}^{n} \alpha_i y_i \nabla x_s k(x_i, x_s) \right|_2^2 \\
+ \frac{1}{n} \sum_{s=1}^{n} \left| -\frac{1}{\rho} \sum_{i=1}^{n} (\alpha_i + \beta_i - C_1) \nabla y_s (v_i^T D^* T D^*) v_s \right|_2^2 \\
= \frac{1}{n} \sum_{s=1}^{n} (\alpha \circ y)^T H_s (\alpha \circ y) \\
+ \frac{1}{\rho^2} (\alpha + \beta - C_1)^T V^T (D^* T D^*)^T (D^* T D^*) V (\alpha + \beta - C_1),
$$

where $k(x_i, x_j) = e^{-\gamma ||x_i - x_j||_2^2}$ is the Gaussian kernel, $h(x_i, x_j, x_s) = \nabla x_s k(x_i, x_s)$, $\nabla x_s k(x_j, x_s) = 4\gamma^2 (x_s - x_i, x_s - x_j) e^{-\gamma ||x_i - x_s||_2^2} e^{-\gamma ||x_j - x_s||_2^2}$, and $H_s$ is a matrix with each element being $H_{s, ij} = h(x_i, x_j, x_s) \in \mathbb{R}^{n \times n}$. Finally, the problem of minimizing $\Gamma(\tau_x, \tau_v)$ is translated into the optimization problem as follows,

$$
\min_{\tau_x \geq 0} \min_{\alpha, \beta} \Theta(f, \zeta).
$$

4.3. Objective function

The proposed Robust SVM+ algorithm aims to solve both optimization problems in Eq. (19) and Eq. (22). Therefore, the whole objective of Robust
SVM+ can be defined as follows

\[
\begin{align*}
\min_{v_j \geq 0, j = \ell + 1, \ldots, n} & \quad \max_{\alpha, \beta} \ 1^T \alpha - \frac{1}{2} (\alpha \circ y)^T K(\alpha \circ y) + C_2 \|1^T \tilde{V} - 1_n^T\|^2_2 \\
& - \frac{1}{2\rho} (\alpha + \beta - C_1)^T V^T D^* D^* V (\alpha + \beta - C_1) + (-1)^\varpi \lambda \Theta(f, \zeta),
\end{align*}
\]

s.t.

\[
\begin{align*}
1^T (\alpha + \beta - C_1) &= 0, \\
y^T \alpha &= 0, \\
0 &\leq \alpha \leq C_3, \\
\beta &\geq 0.
\end{align*}
\]

(23)

where \(\varpi\) is equal to 1 for the optimization of \(\{\alpha, \beta\}\), while \(\varpi\) is equal to -1 for the optimization of \(v\).

5. Optimization

To solve the optimization problem in Eq. (23), we use a two-step alternating optimization strategy which solves the optimization problem with respect to \((\alpha, \beta)\) and \(v\) iteratively. The first step is to solve Eq. (23) with respect to \((\alpha, \beta)\) while fixing \(v\), and the second step is to update \(v\) while fixing \((\alpha, \beta)\). Although this does not guarantee that the solution converges to the global optimum, it is ensured that the objective value obtained at the current iteration will be not larger than the previous one and the solution will finally converges to a good local optimum.

5.1. Optimization of \(\alpha\) and \(\beta\)

When \(v_j\) is fixed, the optimization problem related to \(\alpha\) and \(\beta\) can be written as:
\[
\min_{\alpha, \beta} \frac{1}{2}(\alpha \circ y)^T (K + \frac{2\lambda}{n} \sum_{s=1}^{n} H_s)(\alpha \circ y) - 1_n^T \alpha \\
+ \frac{1}{2\rho}(\alpha + \beta - C_1)^T V^T (D^* D^* + \frac{2\lambda}{\rho} D^* D^* D^* D^*) V (\alpha + \beta - C_1) \\
s.t. \quad 1^T (\alpha + \beta - C_1) = 0, \\
\quad y^T \alpha = 0, \\
\quad 0 \leq \alpha \leq C_3, \\
\quad \beta \geq 0.
\]

(24)

In order to simplify the formulation and make it easy to optimize, we define two matrices \(A = K + \frac{2\lambda}{n} \sum_{s=1}^{n} H_s\), and \(B = V^T (D^* D^* + \frac{2\lambda}{\rho} D^* D^* D^* D^*) V\).

By further defining \(\mu = [(\alpha, \beta)^T, (1 + \frac{1}{\rho} BC_1)^T, (\frac{1}{\rho} BC_1)^T]^T \in \mathbb{R}^{2n}\), and \(G = \begin{bmatrix} A \circ (yy^T) + \frac{1}{\rho} B & \frac{1}{\rho} B \\ \frac{1}{\rho} B & \frac{1}{\rho} B \end{bmatrix} \in \mathbb{R}^{2n \times 2n}\), the optimization problem can be rewritten as

\[
\min_{\mu} \frac{1}{2} \mu^T G \mu - \vartheta^T \mu \\
s.t. \quad 1^T (\alpha + \beta - C_1) = 0, \\
\quad y^T \alpha = 0, \\
\quad 0 \leq \alpha \leq C_3, \\
\quad \beta \geq 0.
\]

(25)

Eq. (25) can be efficiently solved by off-the-shelf quadratic programming solvers.

5.2. Optimization of \(\tilde{V}\)

When \(\alpha\) and \(\beta\) are fixed, since \(\{v_i\}_{i=1}^{\ell}\) are already known, we only solve the optimization problem with respect to \(\tilde{V} = [v_{\ell+1}, \ldots, v_n]\). By defining \(\tilde{\alpha} = [\alpha_{\ell+1}, \ldots, \alpha_n]^T \in \mathbb{R}^{n-\ell}\), \(\tilde{\beta} = [\beta_{\ell+1}, \ldots, \beta_n]^T \in \mathbb{R}^{n-\ell}\), and \(\tilde{C}_1 = [C_1, \ldots, C_1]^T \in \mathbb{R}^{n-\ell}\), the optimization with respect to \(\tilde{V}\) is

\[
\min_{\tilde{V} \succeq 0} \frac{1}{2\rho} (\tilde{\alpha} + \tilde{\beta} - \tilde{C}_1)^T \tilde{V}^T Q \tilde{V} (\tilde{\alpha} + \tilde{\beta} - \tilde{C}_1) + C_2 ||1_{\ell}^T \tilde{V} - 1_{n-\ell}^T||_2^2.
\]

(26)
where \( Q = (D^* D^* - \frac{2\lambda}{\rho} D^* D^* D^* D^*) \in \mathbb{R}^{\ell \times \ell} \). We use the alternating direction method of multipliers (ADMM) technique \[36\] to solve the optimization problem. In order to apply ADMM on this problem, we reformulate Eq. (26) into the following form,

\[
\min_{Z, R, \tilde{V}} - \frac{1}{2\rho} Z^T Q Z + C_2 \|1_{\ell}^T \tilde{V} - 1_n \|_2^2,
\]

s.t. \( Z = R(\bar{\alpha} + \bar{\beta} - \bar{C}_1) \),
\( \tilde{V} = R, \quad \tilde{V} \geq 0 \).

(27)

According to ADMM, the augmented Lagrangian function of Eq. (27) can be written as

\[
L(Z, R, \tilde{V}) = -\frac{1}{2\rho} Z^T Q Z + C_2 \|1_{\ell}^T \tilde{V} - 1_n \|_2^2 + \langle u_1, Z - R(\bar{\alpha} + \bar{\beta} - \bar{C}_1) \rangle + \frac{\theta}{2} \|Z - R(\bar{\alpha} + \bar{\beta} - \bar{C}_1)\|_2^2 + \langle U_2, \tilde{V} - R \rangle + \frac{\theta}{2} \|\tilde{V} - R\|_F^2,
\]

s.t. \( \tilde{V} \geq 0 \).

(28)

where \( u_1 \) and \( U_2 \) are Lagrangian multipliers, and \( \theta > 0 \). According to ADMM, the optimization procedure in Eq. (28) consists of the following alternative iterations:

\[
Z^{(k+1)} = \arg \min_Z L(Z, R^k, \tilde{V}^k)
\]

(29)

\[
R^{(k+1)} = \arg \min_R L(Z^{k+1}, R, \tilde{V}^k)
\]

(30)

\[
\tilde{V}^{(k+1)} = \arg \min_{\tilde{V} \geq 0} L(Z^{k+1}, R^{k+1}, \tilde{V})
\]

(31)

where \( k \) is the iteration number. The augmented Lagrangian is minimized jointly with respect to these three primal variables \( Z, R \) and \( \tilde{V} \). These three primal variables are updated in a sequential and alternating fashion until convergence is achieved.
Algorithm 1 Robust SVM+ for Imperfect Privileged Information

**Input:** Dataset \( \{(x_1, x_1^*, y_1), \ldots, (x_t, x_t^*, y_t), (x_{t+1}, y_{t+1}), \ldots, (x_n, y_n)\} \), parameters \( C_1 > 0, C_2 > 0, \lambda > 0, \rho > 0, \sigma' > 0 \) and \( \theta > 0 \)

1. **Initialization:** \( Z = u_1 = U_2 = 0, \tilde{V} = \frac{1}{\tau}1_{\ell \times (n-\ell)}, \) and \( R = 1_{\ell \times (n-\ell)} \)

2. **repeat**
   
   3. (a) Solve \( \alpha \) and \( \beta \) in Eq. (25) using quadratic programming solvers
   
   4. (b) Solve \( \tilde{V} \) using ADMM technique

   5. **while** not converged **do**

   6. Update \( Z^{(j+1)} \) via Eq. (32)
   
   7. Update \( R^{(j+1)} \) via Eq. (33)
   
   8. Update \( \tilde{V}^{(j+1)} \) via Eq. (34)

   9. Update \( u_1^{(j+1)} \) and \( U_2^{(j+1)} \) via Eq. (35)

   10. Update \( \theta := a \theta \) (\( a > 1 \))

5. **end while**

12. **until** convergence or maximum iteration is reached

**Output:** \( \alpha \) and the final decision function

5.2.1. \( Z \)-subproblem

When solving \( Z \), other variables \( R \) and \( \tilde{V} \) are fixed. The closed-form solution subproblem with respect to \( Z \) in Eq. (29) can be obtained

\[
Z^{(k+1)} = \arg \min_Z \left( -\frac{1}{2\rho} Z^T Q Z + \langle u_1^{(k)}, Z - R^{(k)}(\bar{\alpha} + \bar{\beta} - \bar{C}_1) \rangle \\
+ \frac{\theta}{2} \| Z - R^{(k)}(\bar{\alpha} + \bar{\beta} - \bar{C}_1) \|^2_2 \right)
= \left[ -\frac{1}{2\rho}(Q + Q^T) + \theta I_\ell \right]^{-1}[\theta R^{(k)}(\bar{\alpha} + \bar{\beta} - \bar{C}_1) - u_1^{(k)}].
\]

5.2.2. \( R \)-subproblem

For a fixed \( Z \) and \( \tilde{V} \), the subproblem in Eq. (30) with respect to \( R \) admits a closed-form solution
\[
R^{(k+1)} = \arg \min_{\mathbf{R}} (\mathbf{u}_1^{(k)}, \mathbf{Z}^{(k+1)} - \mathbf{R}(\bar{\alpha} + \bar{\beta} - \bar{C}_1)) + \langle \mathbf{U}_2^{(k)}, \mathbf{V}^{(k)} - \mathbf{R} \rangle + \frac{\theta}{2} \| \mathbf{Z}^{(k+1)} - \mathbf{R}(\bar{\alpha} + \bar{\beta} - \bar{C}_1) \|^2_F + \frac{\theta}{2} \| \mathbf{V}^{(k)} - \mathbf{R} \|^2_F
\]

\[
= \left[ (\mathbf{u}_1^{(k)} + \theta \mathbf{Z}^{(k+1)})(\bar{\alpha} + \bar{\beta} - \bar{C}_1)^T + \theta \mathbf{V}^{(k)} + \mathbf{U}_2^{(k)} \right] \mathbf{T}_R.
\]

where \( \mathbf{T}_R \equiv \left\{ \theta \mathbf{I}_{n-l} + \theta (\bar{\alpha} + \bar{\beta} - \bar{C}_1)(\bar{\alpha} + \bar{\beta} - \bar{C}_1)^T \right\}^{-1} \).

5.2.3. \( \mathbf{V} \)-subproblem

When \( \mathbf{Z} \) and \( \mathbf{R} \) are fixed, by defining a proximal operator \( \mathcal{I}_+ (\mathbf{V}) \) which is the indicator function of the non-negative orthant associated with \( \mathbf{V} \), and the subproblem in Eq. (31) with respect to \( \mathbf{V} \) has the closed-form solution

\[
\mathbf{V} = \arg \min_{\mathbf{V}} C_2 \| \mathbf{I}_l \mathbf{V} - \mathbf{I}_n \mathbf{T} \|_2^2 + \langle \mathbf{U}_2^{(k)}, \mathbf{V} - \mathbf{R}^{(k+1)} \rangle + \frac{\theta}{2} \| \mathbf{V} - \mathbf{R}^{(k+1)} \|^2_F + \mathcal{I}_+ (\mathbf{V})
\]

\[
= \left\{ (2C_2 \mathbf{1}_l \mathbf{1}_n - \theta \mathbf{I}_n)^{-1}(\theta \mathbf{R}^{(k+1)} - \mathbf{U}_2^{(k)}) + 2C_2 \mathbf{1}_l \mathbf{1}_n \mathbf{T} \right\}^+.
\]

where the operator \( \{ \mathbf{a} \}_+ \) turns the negative elements in \( \mathbf{a} \) into 0 and keeps the non-negative elements unchanged.

5.2.4. Lagrangian multipliers

At the end of the \((k+1)\)-th iteration, the lagrangian multipliers \( \mathbf{u}_1 \) and \( \mathbf{U}_2 \) are updated as:

\[
\mathbf{u}_1^{(k+1)} = \mathbf{u}_1^{(k)} + \theta \mathbf{Z}^{(k+1)} - \mathbf{R}^{(k+1)}(\bar{\alpha} + \bar{\beta} - \bar{C}_1),
\]

\[
\mathbf{U}_2^{(k+1)} = \mathbf{U}_2^{(k)} + \theta (\mathbf{V}^{(k)} - \mathbf{R}^{(k+1)}).
\]

6. Experimental evaluation

In this section, we evaluate the performance of the proposed methods and competing methods in three different tasks, including digit image classification, face pose classification and human activity recognition, on three real-world datasets, i.e., the MNIST+ dataset \([7]\), the RGB-D Face dataset \([37]\), and the
Human Activity Recognition dataset [38], respectively. We present the experimental results that evaluate the following competing methods and the proposed methods.

- **SVM**: The standard support vector machine (SVM) that only uses example feature data but does not utilize privileged information.

- **RSVM-RHHQ**: A robust SVM based on the rescaled hinge loss function [39] only using example feature data and without using privileged information.

- **SVM+**: A typical LUPI method [7] using complete privileged information which cannot be applied in the incomplete privileged information situation.

- **L2-SVM+**: An L2-loss SVM+ method [23] based on the $\rho$-SVM formulation which requires complete privileged information.

- **E-SVM+**: An extension of SVM+ [7] which is directly applied for the incomplete privileged information situation that a fraction of training examples are without privileged information.

- **R-SVM+**: The proposed Robust SVM+ method which considers noisy data but uses complete privileged information that all training examples are supplied with privileged information.

- **Incom-SVM+**: The proposed Incomplete SVM+ only deals with the incomplete privileged information.

- **R-Incom-SVM+**: The proposed Robust SVM+ method considering both noisy data and imperfect privileged information.

In the experiments, the results of SVM+, L2-SVM+ and R-SVM+ are used as references of complete privileged information that all training examples are supplied with privileged information, since they cannot be directly applied to the incomplete privileged information situation. For SVM-based methods (SVM
and RSVM-RHHQ) and the SVM+-based methods, the trade-off parameter $C$ is selected in range of $10^{(-2,-1,0,1,2)}$, and the Gaussian kernel is used as the kernel function. The parameter $\rho$ is selected from $10^{(-2,-1,0,1,2)}$ for SVM+-based methods. The scaling parameter $\eta$ for RSVM-RHHQ is selected from $\{0.01, 0.1, 0.5, 1, 2, 3, 10, 100\}$. For R-SVM+, the parameter $\sigma'$ is varied in range of $\{5, 10, 50, 100\}$ and $\lambda$ in range of $10^{(-5,-4,...,0,1)}$. For the proposed Incom-SVM+ and R-Incom-SVM+ methods, the parameter $\sigma'$ is set to 50 and $\lambda$ is varied from $10^{(-5,-3,-1,1)}$. The averaged classification accuracies and standard deviations over 10 times on the test set are reported.

In order to simulate the real situations where noises exist in the data, white Gaussian noises with a specific signal-to-noise ratio (SNR) is added to examples in the validation set and the test set for each dataset. Moreover, for methods that deal with the incomplete privileged information, we randomly select 50% training examples whose privileged information have been supplied, and the remaining 50% training examples have no privileged information.

### 6.1. MNIST+ dataset

The MNIST+ dataset \cite{7} contains 2,943 images of digit “5” and 3,025 images of digit “8” from the MNIST database \cite{40}. The size of each original image is 28×28. In order to make it more difficult, all the images of two digits are resized into 10×10 pixels in the MNIST+ dataset. Each image is created with a holistic (poetic) description which is translated by experts into a 21-dimensional feature vector. The 100-dimensional vectors of raw pixels of each image are used as the example feature data, while its corresponding 21-dimensional feature vector descriptions are used as the privileged feature data.

In the experiment, we randomly select 90 examples of 10×10 images and their corresponding privileged feature data as the training set, 4,002 examples as the validation set, and remaining 1,866 examples as the test set. In detail, for E-SVM+, Incom-SVM+ and R-Incom-SVM+ methods which deal with incomplete privileged information, only 45 training examples are supplied with privileged information. While for SVM+, L2-SVM+ and R-SVM+, all training examples
are provided with privileged information.

Table 1 reports the classification accuracies of the proposed methods and competing methods with different SNRs of 4, 6, 8, 10 and 15. We can observe that performances of all methods are affected by different SNRs. When SNR is smaller, E-SVM+ and SVM+ shows smaller superiority than SVM and RSVM-RHHQ, and even L2-SVM+ gets worse performance than SVM and RSVM-RHHQ. This may be due to their lack of clear strategies to handle noise and imperfect privileged information. While the proposed methods obviously outperform other competing methods with varying SNRs, which demonstrates the effectiveness and robustness of the proposed methods. When SNRs get larger, all LUPI methods get better results than SVM and RSVM-RHHQ, this shows the effectiveness of using privileged information. Note that LUPI methods using incomplete privileged information even outperform their corresponding methods using complete privileged information. It may be because on this dataset, not all the provided privileged information is accurate and helpful for the improvement of performance when noise exist in example data. And the re-constructed missing privileged information via the proposed Incom-SVM+ and Incom-R-SVM+ methods are better to learn a model. Moreover, R-Incom-SVM+ gets better results than Incom-SVM+, which shows that R-Incom-SVM+ is more robust to noises and possibly inaccurate estimations of missing privileged information. We report the performances of R-SVM+ in Figure 1 to discuss the

<table>
<thead>
<tr>
<th>SNR</th>
<th>SVM</th>
<th>RSVM-RHHQ</th>
<th>E-SVM+</th>
<th>Incom-SVM+</th>
<th>R-Incom-SVM+</th>
<th>SVM+</th>
<th>L2-SVM+</th>
<th>R-SVM+</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>82.36±1.75</td>
<td>83.29±1.10</td>
<td>83.64±1.62</td>
<td>84.87±1.97</td>
<td><strong>85.81±1.49</strong></td>
<td>82.99±2.30</td>
<td>80.70±2.45</td>
<td>84.60±1.54</td>
</tr>
<tr>
<td>6</td>
<td>83.54±1.55</td>
<td>83.41±0.86</td>
<td>84.73±1.75</td>
<td>86.65±2.04</td>
<td><strong>87.03±1.65</strong></td>
<td>84.41±2.56</td>
<td>82.00±2.78</td>
<td>86.38±1.59</td>
</tr>
<tr>
<td>8</td>
<td>84.74±1.60</td>
<td>85.28±0.78</td>
<td>85.58±1.75</td>
<td>88.08±1.04</td>
<td><strong>88.45±1.52</strong></td>
<td>86.05±2.10</td>
<td>84.60±1.91</td>
<td>87.71±1.65</td>
</tr>
<tr>
<td>10</td>
<td>86.12±1.38</td>
<td>85.74±0.80</td>
<td>87.42±1.26</td>
<td>89.39±1.13</td>
<td><strong>89.64±1.49</strong></td>
<td>87.55±1.70</td>
<td>85.94±1.65</td>
<td>89.03±1.31</td>
</tr>
<tr>
<td>15</td>
<td>87.78±1.18</td>
<td>86.23±0.92</td>
<td>89.49±1.26</td>
<td>90.87±1.06</td>
<td><strong>91.34±1.20</strong></td>
<td>90.03±1.10</td>
<td>89.28±0.97</td>
<td>90.95±1.08</td>
</tr>
</tbody>
</table>

Table 1: Classification accuracies (mean ± standard deviation, %) on the MNIST+ dataset. The best results on each row are in boldface.
R-SVM+ outperforms other methods with varying SNRs, which demonstrates that the necessity and effectiveness of considering both components. And generally R-SVM+ without $\epsilon$ plays a more important role on the performance of R-SVM+ in these cases.

We further evaluate the robustness of the proposed Incom-SVM+, R-Incom-SVM+, R-SVM+ algorithm and other compared SVM+-based methods, when privileged information (PF) are also very noisy, by adding Gaussian noises to the PF examples with SNRs of 4, 6, 8, 10, and 15. We can observe in Figure 2 (a) that R-Incom-SVM+ and R-SVM+ are more robust to noise and obviously outperform other methods. Differing from the results shown in Table 1, R-SVM+ gets the best results and outperforms Incom-SVM+ and R-Incom-SVM+. This may be because when available privileged information is more noisy, the re-constructed incomplete privileged information may also be in worse quality than original complete privileged information. In detail, when SNRs are equal to 4, 6, 8, 10, and 15, R-SVM+ obtains gains in accuracy of $+9.79\%$, $+8.96\%$, $+6.61\%$, $+5.77\%$, and $+2.71\%$ over SVM+, $+2.18\%$, $+2.18\%$, $+2.97\%$, $+2.64\%$, and $+1.72\%$ over E-SVM+, and $+4.28\%$, $+4.75\%$, $+3.84\%$, $+3.31\%$, and $+1.33\%$ over L2-SVM+, respectively. We also analyze the effects of the percentages of available privileged information on the performance of E-SVM+, the proposed Incom-SVM+ and the proposed R-Incom-SVM+ in
Figure 2: Discussions of performances of different methods.

(a) Influence of noisy privileged information with varying SNRs. (b) Effects of different percentages of available privileged information.

Figure 2 (b). It is observed that the accuracies of these methods are influenced by different percentages. And Incom-SVM+ and R-Incom-SVM+ outperform E-SVM+ in all cases. When the percentage is close to 0.5, Incom-SVM+ and R-Incom-SVM+ get relatively higher accuracies. This may be because when the percentage is too small, the number of available privileged information is too small for Incom-SVM+ and R-Incom-SVM+ to re-construct incomplete privileged information in good-quality. When the percentage gets relatively larger, only few privileged information needs to be re-construct and the performance is affected by the complete privileged information. Figure 3 shows the convergence of the proposed R-Incom-SVM+ method, using the MNIST+ dataset as an example. Figures 3 (a) and (b) depict the value of the objective in Eq. (23) when $\varpi = 1$ and $\varpi = -1$, respectively, as the number of iterations increases. According to the convergence curves of the objective function values, the proposed R-Incom-SVM+ method can converge to a stable value after a number of iterations.

6.2. RGB-D Face dataset

The RGB-D Face dataset [37] contains 1,581 RGB images and their depth counterparts which are taken from 31 people in different face poses and expressions using a Kinect sensor by repeating 3 times for each person. Since
the number of images for each face pose and expression is relatively small, we merge all categories into four main groups including the pose of looking up, the pose of looking forward, the pose of looking down and the pose of having facial expressions.

In the experiment, the RGB images are used as example feature data, and the depth images are served as privileged information. We focus on binary classification on each pair of groups. The dataset is split into a training set of 40% RGB-depth image pairs per class, a validation set of 30% image pairs per class, and a test set of 30% image pairs per class. Each RGB and its corresponding depth image are cropped into a size of 150×150 pixels and each RGB image is converted into a gray image. We extract LBP features for each non-overlapping 15×15 subregion from each image, and then concatenate LBP features extracted from all subregions. Since the dimension of the derived features is high, we perform PCA to derive a 150-dimensional compact representation.

For methods E-SVM+ and the proposed Incom-SVM+ and R-Incom-SVM+ methods, they solve the incomplete privileged information problem, where 50% examples in the training set are supplied with privileged information, while the rest training examples are not with privileged information. In order to evaluate the performance of all methods when noises exist, we add Gaussian noises with an SNR of 8 to examples in the validation and test set.
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.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>SVM-RSRQ</th>
<th>SVM+</th>
<th>Income-SVM+</th>
<th>R-Income-SVM+</th>
<th>SVM+</th>
<th>L2-SVM+</th>
<th>R-SVM+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up vs Forward</td>
<td>95.91 ± 1.38</td>
<td>96.47 ± 1.06</td>
<td>96.64 ± 1.05</td>
<td>96.79 ± 1.34</td>
<td>96.87 ± 1.95</td>
<td>96.59 ± 1.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up vs Down</td>
<td>95.91 ± 1.38</td>
<td>96.47 ± 1.06</td>
<td>96.64 ± 1.05</td>
<td>96.79 ± 1.34</td>
<td>96.87 ± 1.95</td>
<td>96.59 ± 1.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up vs Expression</td>
<td>96.83 ± 1.80</td>
<td>97.15 ± 1.75</td>
<td>97.41 ± 1.34</td>
<td>97.58 ± 1.59</td>
<td>97.71 ± 1.51</td>
<td>97.72 ± 1.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward vs Down</td>
<td>99.58 ± 1.09</td>
<td>99.60 ± 1.01</td>
<td>99.62 ± 1.03</td>
<td>99.63 ± 1.07</td>
<td>99.64 ± 1.07</td>
<td>99.63 ± 1.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Down vs Expression</td>
<td>95.51 ± 2.11</td>
<td>96.88 ± 1.80</td>
<td>96.88 ± 1.80</td>
<td>96.88 ± 1.80</td>
<td>96.88 ± 1.80</td>
<td>96.88 ± 1.80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The performances of different methods on the RGB-D Face dataset are reported in Table 2. In general, it is observed that SVM+-based methods get better results than the SVM-based methods. This is an intuitive result because when a person makes different poses like looking forward and looking down, their depth images are quite different. This can help LUPI models learn better with depth information.

The proposed Incom-SVM+ achieve better performances in five cases than E-SVM+, and and R-Incom-SVM+ gets better results than E-SVM+ in all six cases. They show that the strategy of the proposed methods to construct missing privileged information is effective. Moreover, R-Incom-SVM+ is superior to Incom-SVM+ in all the six cases, and this shows that R-Incom-SVM+ has better robustness against data noise and possible errors in estimated missing privileged information. R-Incom-SVM+ also shows its superiority to SVM+ and L2-SVM+ in all six cases. This shows the effectiveness and robustness of the proposed R-incom-SVM+.

6.3. Human Activity Recognition dataset

The Human Activity Recognition dataset [38] is created to gather context information about people actions, by sensing six different human activities of 30 people using a waist-mounted Samsung Galaxy S II smartphone with inertial sensors embedded. The six human activities include walking, walking upstairs, walking downstairs, sitting, standing, and laying down. The dataset contains
10,299 examples obtained by gyroscopes and smartphone accelerometers and so on.

In the experiment, we consider binary classification of each activity category and the rest of activity categories. The 200 examples from the desired category and 200 examples from the remaining categories are randomly split into the training set. The 600 examples from the desired category and 600 examples from the remaining categories are randomly selected for testing. The remaining examples from the desired category and the same number of examples from the rest of categories are split into the validation set. The acceleration raw signals and their additional time raw signals are used as example features, and angular velocity raw signals from the gyroscope and their additional time raw signals are used as privileged information. In the experiment, we add Gaussian noises with an SNR of 4 to the validation and test examples.

The classification accuracies of different methods on the Human Activity Recognition dataset are reported in Table 3. Generally, the proposed R-SVM+ method get relatively better results than other methods. Especially in the “standing” case, R-SVM+ obtains gains in accuracy of +2.3%, +4.4%, +4.2%, +5.1%, +6.6% and +9.2% over SVM, RSVM-RHHQ, E-SVM+, Incom-SVM+, SVM+, and L2-SVM+. Compared with SVM, SVM+ and L2-SVM+ get obviously worse results in “sitting” and “standing” cases, this shows that when noises exist, the performances of these methods using privileged information are easy to be affected. While R-SVM+ can still get relatively good outperforms in all cases, which shows its effectiveness of improving the model’s robustness against noises. Moreover, R-Incom-SVM+ performs better than E-SVM+ and

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.
Methods & No Privileged Information & Incomplete Privileged Information & Complete Privileged Information &
SVM & RSVM-RHRQ & E-SVM+ & R-Incom+SVM+ & SVM+ & L2-SVM+ & R-SVM+
Accuracy & 63.82±0.46 & 60.14±0.40 & 60.35±0.98 & 62.67±0.77 & 64.92±0.20 & 64.55±0.20 & 67.17±0.18

Table 4: Classification accuracies (mean ± standard deviation, %) on the Tiny-ImageNet dataset with an SNR of 4. Best accuracies are highlighted in boldface.

Incom-SVM+ in all cases, especially in the “Walking upstairs”, “Sitting” and “Standing” cases, which suggests that the performance of Incom-SVM+ is obviously influenced by noises and the consideration of the robustness of the Incom-SVM+ model is helpful.

6.4. Tiny ImageNet dataset

The Tiny ImageNet dataset is a subset of the ILSVRC-2012 classification dataset. It consists of 200 object classes, and for each object class it provides 500 training images, 50 validation images, and 50 test images. All images have been down-sampled to 64×64×3 pixels. The training and validation sets are released with images and annotations, including both class labels and bounding boxes. In the experiment, the released 500 training images for each class are used as the training set, and the released 50 validation images for each class are used as test set. All 64×64×3 pixel images are used as example feature data and the image portions in the bounding boxes are used as privileged information. Moreover, we extract deep features of RGB images and the image portions in the bounding boxes with a fine-tuned ResNet18 model using pre-trained weight from ImageNet. For fine tuning, since the size of input image is smaller than original 224×224×3 images from ImageNet, for the first convolution layer we set the kernel size, stride and padding to 3x3, 1, and 1, respectively, and remove the max pool layer. Finally, the output of the average pool layer for each image is a 512-dimension features which is used in the experiment.

In the experiment, we use one-versus-rest techniques to solve the classification problem and build a binary classification model for each object class. For training, the 500 examples from the desired class and 500 examples randomly selected from the remaining classes are used. The test set is fixed containing
10,000 examples. In the experiment, we add Gaussian noises with an SNR of 4 to the test examples. The overall classification accuracies are reported in Table 4. It is observed that SVM+, L2-SVM+ and the proposed R-SVM+ get relatively better results than other compared methods, which demonstrate that using the image portion in the bounding box as complete privileged information is helpful for the classification task. The proposed R-SVM+ outperforms all compared methods, and this shows the effectiveness of improving the robustness of the model against noises. R-Incom-SVM+ performs better than E-SVM+, this suggests that the strategy of both re-constructing incomplete privileged information and improving the robustness is effective. Moreover, R-Incom-SVM+ gets a little lower accuracy than SVM, which may be because that when building a binary classification model for each class and remaining 199 classes, the privileged information from so many remaining classes are very different and complex and may not be easy to re-construct perfectly.

7. Conclusion

This paper proposes a Robust SVM+ method to deal with noise in the example data and imperfect privileged information. By theoretically analyzing the lower bound of the perturbations that may influence the judgement of SVM+, the derived Robust SVM+ can be more robust to the noises of example data and inaccurate or noisy privileged information. In order to deal with the incomplete privileged information situation, we propose an Incomplete SVM+ which learns and re-constructs the incomplete privileged information using the available privileged information. And we further extend Robust SVM+ to the framework of Incomplete SVM+ to deal with imperfect privileged information. Experiment results on different real-world datasets demonstrate that the proposed Robust SVM+ method is effective compared with different competing methods.
8. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61822113, 61771349, and Australian Research Council Projects: FL-170100117, DP-180103424, DP-140102164, and LP-150100671, DE-180101438.

References


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


