Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Multiclass Capped ℓ_p-Norm SVM for Robust Classifications

Feiping Nie,[1,2] **Xiaoqian Wang,**[2] **Heng Huang**[2*]
[1]School of Computer Science, OPTIMAL, Northwestern Polytechnical University, Xian 710072, Shaanxi, P. R. China
[2]Department of Computer Science and Engineering, University of Texas at Arlington, USA
feipingnie@gmail.com, xqwang1991@gmail.com, heng@uta.edu

Abstract

Support vector machine (SVM) model is one of most successful machine learning methods and has been successfully applied to solve numerous real-world application. Because the SVM methods use the hinge loss or squared hinge loss functions for classifications, they usually outperform other classification approaches, *e.g.* the least square loss function based methods. However, like most supervised learning algorithms, they learn classifiers based on the labeled data in training set without specific strategy to deal with the noise data. In many real-world applications, we often have data outliers in train set, which could misguide the classifiers learning, such that the classification performance is suboptimal. To address this problem, we proposed a novel capped ℓ_p-norm SVM classification model by utilizing the capped ℓ_p-norm based hinge loss in the objective which can deal with both light and heavy outliers. We utilize the new formulation to naturally build the multiclass capped ℓ_p-norm SVM. More importantly, we derive a novel optimization algorithms to efficiently minimize the capped ℓ_p-norm based objectives, and also rigorously prove the convergence of proposed algorithms. We present experimental results showing that employing the new capped ℓ_p-norm SVM method can consistently improve the classification performance, especially in the cases when the data noise level increases.

Introduction

As one of the most fundamental problems in data mining, classification has numerous applications in different areas such as information retrieval (Cao et al. 2009; Sriram et al. 2010), computer vision (Krizhevsky, Sutskever, and Hinton 2012), bioinformatics (Brown et al. 2000), medical image computing (Chen, Daponte, and Fox 1989), natural language processing (Wang and Mannin 2012) *etc*. Given training data from multiple classes, the classification task is to learn the classifiers in a supervised way and find the correct class to which a test example belongs. Many classification models have been proposed in literature. Among them, the support vector machine (SVM) (Boser, Guyon, and Vapnik 1992) is one of the most successful classification models and has been applied to solve various applications. One of main reasons for SVM models (Mangasarian 2002; Keerthi and DeCoste 2005; Lin, Weng, and Keerthi 2008; Chang, Hsieh, and Lin 2008; Hsieh et al. 2008) to outperform other classification methods is their unilateral loss function, *e.g.* hinge loss or squared hinge loss. The unilateral loss is more suitable for classification tasks than the bilateral loss, which has been used in regression models.

However, like most supervised learning algorithms, existing SVM models learn classifiers based on the labeled data in training set without considering the noise problem. In many real-world applications, we often have data outliers in train set, *e.g.* the incorrectly labeled data, the data significantly different to other data in the same class, *etc*. These data outliers could mislead the classifiers training task, such that the learned classifiers are not optimal and the classification performance is reduced. Thus, the robust classification model is desired to deal with the classification tasks in real-world applications. Although sparse learning models have been applied to SVM methods in literature, such as ℓ_1-SVM (Bradley and Mangasarian 1998), $\ell_{2,1}$-SVM (Cai et al. 2011), Hybrid Huberized SVM (Wang, Zhu, and Zou 2007), Sparse SVM (Cotter, Shalev-Shwartz, and Srebro 2013), these methods mainly focus on selecting significant features or reducing the number of support vectors to improve the classification tasks. None of them are specifically designed to deal with the data outliers for robust classifications.

To address this challenging problem, we propose a novel capped ℓ_p-norm SVM classification model by utilizing the capped ℓ_p-norm based loss function. Different to the hinge loss and squared hinge loss used in existing SVM methods, which are not robust to data outliers, the capped ℓ_p-norm is theoretically robust to both light and heavy outliers. Because the data outliers usually have large residues, the capped norm can help the model eliminate these outliers in model training process. In term of multiclass classifications, the existing research often uses the one-vs-one or one-vs-rest strategies to utilize the binary SVM classifier for solving multiclass classification tasks, but the label ambiguity problem has been well-known for such situations. Thus, we also introduce the new formulation for multiclass SVM model, which can directly solve the multiclass classification problem. Based on the new multiclass SVM formulation, we can

*Corresponding author. This work was partially supported by the following grants: NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, NIH R01 AG049371.

naturally build the multiclass capped ℓ_p-norm SVM model.

The capped ℓ_p-norm in our new loss functions increases the non-smoothness of the objectives, thus it is challenging to solve them. To tackle this problem, we introduce new optimization algorithms to efficiently solve the proposed capped ℓ_p-norm SVM problem. The proposed optimization algorithm can solve the general capped ℓ_p-norm based objectives (even for general concave functions) with rigorously proved local optimum convergence. As a result, our new algorithm can be applied to solve many other capped norm applications, such as capped ℓ_p-norm logistic regression, *etc.* The validation experiments have been conducted on six benchmark datasets. All empirical results demonstrate that our new capped ℓ_p-norm SVM method is robust to data outliers and consistently improve the classification performance. The experimental results also show that our new method is not sensitive to parameters, but other related methods are sensitive to parameter tuning. Thus, our new capped ℓ_p-norm SVM method is suitable for practical applications.

Our main contributions of this paper can be summarized in the following three folds:

1) We propose a novel capped ℓ_p-norm SVM classification model for robust classifications. The capped ℓ_p-norm based loss function is robust to the data outliers by eliminating the abnormal data points with very large residues, such that the classification model training step is robust to the noise and incorrect labels.

2) Because the proposed capped ℓ_p-norm SVM loss function is non-smooth and non-convex, we introduce a new optimization algorithm to solve it. More importantly, our new optimization algorithm can solve the general capped ℓ_p-norm based objectives, thus our optimization algorithm can be applied to other capped norm problems.

3) In order to avoid the ambiguity problem existing in well-known one-vs-one or one-vs-rest multiclass classification strategies, we present the new formulation for multiclass SVM model and further derive the multiclass capped ℓ_p-norm SVM method.

Motivation and Proposed New Objective

In a classification task, given the training data $X = [x_1, \cdots, x_n] \in \Re^{d\times n}$, we usually learn a linear model by solving the following problem:

$$\min_{w,b} \sum_{i=1}^{n} \xi(w, b\,|x_i\,) + \gamma \left\|w\right\|_2^2 , \tag{1}$$

where the first term is the loss and the second term is the regularization, γ is a parameter to balance these two terms. The $w \in \Re^{d\times 1}$ is the projection vector and $b \in \Re$ is the bias term in the linear model $w^T x + b$.

In the past decades, many loss functions have been proposed for learning classifiers. Different loss functions lead to various classifiers. One of the most popular and successful loss function is the hinge loss function, which leads to the classical Support Vector Machines (SVM) classifier.

The hinge loss function is defined as follows:

$$\xi_h(w, b\,|x_i\,) = \max\left(1 - y_i(w^T x_i + b), 0\right) \tag{2}$$

where $y_i \in \{1, -1\}$ is the given class label of the data point x_i. Another popularly used loss function is the squared hinge loss function, which is defined as:

$$\begin{aligned}\xi_{sh}(w, b\,|x_i\,) &= \left(\xi_h(w, b\,|x_i\,)\right)^2 \\ &= \left(\max\left(1 - y_i(w^T x_i + b), 0\right)\right)^2 .\end{aligned} \tag{3}$$

In contrast with the hinge loss function, the squared hinge loss function is differentiable, such that it can be easily optimized.

Unilateral Loss vs Bilateral Loss

In Fig. 1(a) and Fig. 1(c), the hinge loss function and the squared hinge loss function are plotted. Both of them are unilateral loss. In contrast, the bilateral loss corresponding to the hinge loss is the ℓ_1-norm loss as follows (as shown in Fig. 1(b)):

$$\begin{aligned}\xi_{\ell_1}(w, b\,|x_i\,) &= \left|1 - y_i(w^T x_i + b)\right| \\ &= \left|w^T x_i + b - y_i\right| .\end{aligned} \tag{4}$$

The bilateral loss corresponding to the squared hinge loss is the ℓ_2-norm loss (as shown in Fig. 1(d)):

$$\begin{aligned}\xi_{\ell_2}(w, b\,|x_i\,) &= (1 - y_i(w^T x_i + b))^2 \\ &= (w^T x_i + b - y_i)^2 .\end{aligned} \tag{5}$$

For classification tasks, if the data point x_i is correctly classified, *i.e.* $y_i(w^T x_i + b) - 1 \geq 0$, the loss should be zero. The unilateral loss functions (shown in Fig. 1(a)) and Fig. 1(c)) meet this requirement, but the bilateral loss functions (shown in Fig. 1(b)) and Fig. 1(d)) do not. Therefore, the unilateral loss based classification models are more suitable for classification than the bilateral loss based classification models. In other words, **the unilateral loss is more suitable for classification problem while the bilateral loss is more suitable for regression problem.**

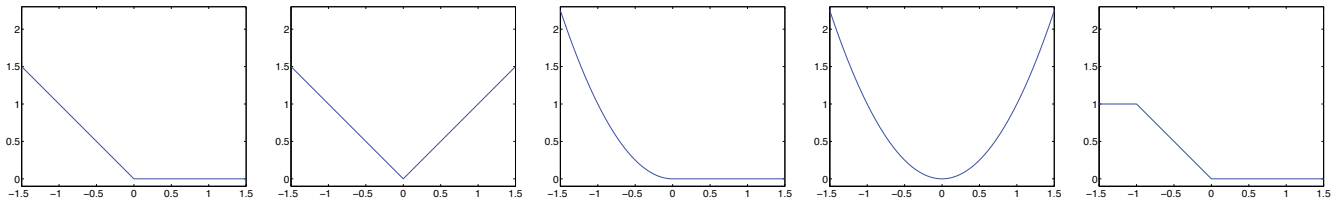
Robust Unilateral Loss with Capped ℓ_p-Norm

From the above analysis we know that the hinge loss and the squared hinge loss are suitable for classification tasks. However, as can been seen in Fig. 1(a) and Fig. 1(c), if the data point x_i is not correctly classified, the loss could be infinite. Therefore, the hinge loss and the squared hinge loss are not robust enough to data outliers.

To solve this problem, in this paper, we propose to use the capped ℓ_p-norm in the unilateral loss for robust classification. In recent research work (Zhang 2008; 2010; Huo, Nie, and Huang 2016; Gao et al. 2015; Jiang, Nie, and Huang 2015), the capped ℓ_1-norm was successfully used to approximate the ℓ_0-norm. Our new robust unilateral loss function using capped ℓ_p-norm is defined as follows:

$$\begin{aligned}\xi_{ch}(w, b\,|x_i\,) &= \min\left(\left(\xi_h(w, b\,|x_i\,)\right)^p, \varepsilon\right) \\ &= \min\left(\left(\max(1 - y_i(w^T x_i + b), 0)\right)^p, \varepsilon\right) .\end{aligned} \tag{6}$$

To illustrate the advantage to utilize the capped ℓ_p-norm in unilateral loss function, we plot capped ℓ_1-norm based hinge



(a) Hinge (b) ℓ_1-norm (c) Squared hinge (d) ℓ_2-norm (e) Capped hinge

Figure 1: The plots of different unilateral loss functions and bilateral loss functions. The abscissa is the value of $y_i(w^T x_i+b)-1$. In term of classification performance, the hinge loss or square hinge loss functions usually are better than the least square loss functions, because of they use one-side functions as visualized. The capped hinge loss function is further robust to data outliers, because the abnormal data points with large residues are eliminated (capped) as illustrated.

loss function[1] in Fig. 1(e), where $p = 1$ and $\varepsilon = 1$. From Fig. 1(e), we can see that no matter how misclassified the data point is, the loss is at most ε. This property makes the loss function very robust to data outliers, especially for the heavy outliers. With this robust unilateral loss, we propose to solve the following capped ℓ_p-norm SVM problem for robust classification:

$$\min_{w,b} \sum_{i=1}^{n} \xi_{ch}(w, b\,|x_i\,) + \gamma \left\|w\right\|_2^2 . \tag{7}$$

Multiclass Capped ℓ_p-Norm SVM

In the above capped ℓ_p-norm SVM problem (7), we only consider the binary class case. In order to extend the problem in (7) to the multiclass case, firstly we need to reformulate the unilateral loss function. It can be easily verified that the hinge loss in Eq. (2) can be reformulated as follows (Xiang et al. 2012):

$$\xi_h(w, b\,|x_i\,) = \min_{m_i \geq 0} \left|w^T x_i + b - y_i - y_i m_i\right|, \tag{8}$$

where the introduced $m_i \in \Re$ is a slack variable to encode the unilateral loss of x_i. Similarly, the proposed robust unilateral loss defined in Eq. (6) can be reformulated as follows:

$$\begin{aligned} &\xi_{ch}(w, b\,|x_i\,) \\ &= \min\left(\min_{m_i \geq 0} \left|w^T x_i + b - y_i - y_i m_i\right|^p, \varepsilon\right). \end{aligned} \tag{9}$$

In the multiclass problem, we are given a class label vector $y_i \in \Re^{c\times 1}$ for each data point x_i, where the k-th element of y_i is 1 if x_i is labeled as class k, and is -1 otherwise. Inspired by Eq. (9) in the binary class case, the robust unilateral loss with ℓ_p-norm in the multiclass case can be naturally defined as follows:

$$\begin{aligned} &\xi_{chm}(W, b\,|x_i\,) \\ &= \min\left(\min_{m_i \geq 0} \left\|W^T x_i + b - y_i - y_i \circ m_i\right\|_2^p, \varepsilon\right), \end{aligned} \tag{10}$$

where $W \in \Re^{d\times c}$ is the projection matrix and $b \in \Re^{c\times 1}$ is the bias term in the linear model $W^T x + b$, the introduced $m_i \in \Re^{c\times 1}$ is a slack variable to encode the unilateral loss of x_i.

Therefore, the capped ℓ_p-norm SVM problem (7) in the binary class case can be extended to the multiclass case as follows:

$$\min_{W,b} \sum_{i=1}^{n} \xi_{chm}(W, b\,|x_i\,) + \gamma \left\|W\right\|_F^2 \tag{11}$$

Based on the definition in Eq. (10), the problem (11) can be written as

$$\begin{aligned} \min_{W,b,M\geq 0} \quad & \sum_{i=1}^{n} \min\left(\left\|W^T x_i + b - y_i - y_i \circ m_i\right\|_2^p, \varepsilon\right) \\ & + \gamma \left\|W\right\|_F^2 \end{aligned} \tag{12}$$

where $M \in \Re^{c\times n}$ with the i-th column as m_i.

The capped hinge loss introduces the difficulty to optimize the problem in (12). For example, for capped hinge loss in Fig. 1(e), the non-smoothness comes from the definitions of both capped norm and ℓ_1-norm, and the standard sparse learning optimization algorithms cannot be directly applied to solve the capped norm based objectives. Although recent research work on capped norm presented certain optimization algorithms (Sun, Xiang, and Ye 2013; Gong, Ye, and Zhang 2013), they mainly solve the bi-convex problems and cannot be directly applied to solve our capped ℓ_p-norm SVM objective. To address this problem, in next section, we are going to introduce an efficient optimization framework, which can be utilized to solve various capped norm based objectives.

Optimization Algorithm for Multiclass Capped ℓ_p-Norm SVM

In this section, we will introduce an efficient and theoretical guaranteed algorithm to solve a general problem with the problem (12) as a special case.

Algorithm to Solve a General Problem

We consider to solve a general problem as follows:

$$\min_{x\in\mathcal{C}} f(x) + \sum_i h_i(g_i(x)), \tag{13}$$

[1] Interestingly, it can be proved that minimizing this loss function ($p = 1$ and $\varepsilon = 1$) is equivalent to minimizing the classification error on training data.

where $h_i(x)$ is an arbitrary **concave** function in the domain of $g_i(x)$, and $x \in \mathcal{C}$ is arbitrary constraint on x. Please note that x and $g_i(x)$ can be scalar, vector or matrix. The details to solve problem (13) is described in Algorithm 1, where $h_i'(g_i(x))$ denotes any supergradient of the concave function h_i at point $g_i(x)$.

Algorithm 1 Re-weighted method to solve problem (13).

Initialize $x \in \mathcal{C}$
while not converge **do**
1. For each i, calculate the supergradient of the concave function: $D_i = h_i'(g_i(x))$
2. Update x by the optimal solution to the problem: $\min_{x \in \mathcal{C}} f(x) + \sum_i Tr(D_i^T g_i(x))$
end while

Convergence Analysis of Algorithm 1

Theorem 1 *The Algorithm 1 will decrease the objective value of the problem (13) in each iteration until it converges.*

Proof: Suppose the updated x is $\tilde{x}$ in the Step 2 of Algorithm 1. According to Step 2, we know:

$$f(\tilde{x}) + \sum_i Tr(D_i^T g_i(\tilde{x})) \leq f(x) + \sum_i Tr(D_i^T g_i(x)), \quad (14)$$

where the equality holds when and only when the algorithm converges.

Because $h_i(x)$ is concave for each i, according to the definition of supergradient, we have:

$$h_i(g_i(\tilde{x})) - h_i(g_i(x)) \leq Tr(D_i^T g_i(\tilde{x})) - Tr(D_i^T g_i(x)) .$$

Thus, we have:

$$\begin{aligned} &\sum_i h_i(g_i(\tilde{x})) - \sum_i Tr(D_i^T g_i(\tilde{x})) \\ &\leq \sum_i h_i(g_i(x)) - \sum_i Tr(D_i^T g_i(x)) . \end{aligned} \quad (15)$$

Summing Eq. (14) and Eq. (15) on both sides, we arrive at:

$$f(\tilde{x}) + \sum_i h_i(g_i(\tilde{x})) \leq f(x) + \sum_i h_i(g_i(x)) . \quad (16)$$

Please note that the equality in Eq. (16) holds only when the algorithm converges. Thus the Algorithm 1 will monotonically decrease the objective of the problem (13) in each iteration until the algorithm converges. □

More importantly, we need further prove the converged solution is a local minimum of the problem (13). To this end, we need the following chain rule:

Lemma 1 (chain rule) *If both x and $g(x)$ are scalar, vector or matrix, we have*

$$\frac{\partial h(g(x))}{\partial x} = \frac{Tr\left((h'(g(x)))^T \partial g(x)\right)}{\partial x} . \quad (17)$$

The convergence analysis of our algorithm is summarized by the following theorem:

Theorem 2 *The Algorithm 1 will converge to a stationary point of the problem (13).*

Proof: The Lagrangian function of the problem (13) is

$$\mathcal{L}_1(x, \lambda) = f(x) + \sum_i h_i(g_i(x)) - r(x, \lambda) , \quad (18)$$

where $r(x, \lambda)$ is a certain function to encode the constraint $x \in \mathcal{C}$ in the Lagrangian function. Based on the KKT condition, by setting the derivative of $\mathcal{L}_1(x, \lambda)$ *w.r.t.* x to zero, we have:

$$\frac{\partial \mathcal{L}_1(x, \lambda)}{\partial x} = f'(x) + \sum_i \frac{\partial h_i(g_i(x))}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} = 0 . \quad (19)$$

According to the chain rule, Eq. (19) can be rewritten as

$$\begin{aligned} &\frac{\partial \mathcal{L}_1(x, \lambda)}{\partial x} \\ &= f'(x) + \sum_i \frac{Tr\left((h_i'(g_i(x)))^T \partial g_i(x)\right)}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} \\ &= 0 . \end{aligned} \quad (20)$$

On the other hand, in the second step of Algorithm 1, we solve the problem $\min_{x \in \mathcal{C}} f(x) + \sum_i Tr(D_i^T g_i(x))$. The Lagrangian function of this problem is:

$$\mathcal{L}_2(x, \lambda) = f(x) + \sum_i Tr(D_i^T g_i(x)) - r(x, \lambda) . \quad (21)$$

By setting the derivative of $\mathcal{L}_2(x, \lambda)$ *w.r.t.* x to zero, we have:

$$\frac{\partial \mathcal{L}_2(x, \lambda)}{\partial x} = f'(x) + \sum_i \frac{\partial Tr(D_i^T g_i(x))}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} = 0 . \quad (22)$$

Thus, we find a solution satisfying Eq. (22) in each iteration according to the second step of Algorithm 1. In the convergence of the Algorithm 1, please note that $D_i = h_i'(g_i(x))$ according to the first step of Algorithm 1, Eq. (22) is equal to:

$$f'(x) + \sum_i \frac{Tr\left((h_i'(g_i(x)))^T \partial g_i(x)\right)}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} = 0 ,$$

which is exactly the same as the KKT condition of the problem (13) in Eq. (20). Therefore, in the convergence of Algorithm 1, the solution x satisfies the KKT condition of the problem (13). Thus the Algorithm 1 will converge to a stationary point of the problem (13), which usually is also a local minimum. □

Empirical evidences show Algorithm 1 converges very fast and usually converges in 20 iterations.

Optimization Algorithm to Solve Problem in (12)

We define the function $h_i(\cdot)$ as follows:

$$h_i(g_i(W, b, m_i)) = \min\left(g_i(W, b, m_i)^{\frac{p}{2}}, \varepsilon\right) , \quad (23)$$

where

$$g_i(W, b, m_i) = \left\| W^T x_i + b - y_i - y_i \circ m_i \right\|_2^2 .$$

It can be easily seen that the function $h_i(\cdot)$ is concave when $0 < p \leq 2$. Therefore, the problem (12) is also a special case of the general problem (13) when $0 < p \leq 2$, and thus we can apply the Algorithm 1 to solve the proposed problem (12) for multiclass robust classification.

According to the second step of Algorithm 1, we need to solve the following problem in each iteration for solving problem (12):

$$\min_{W,b,M\geq 0} \sum_{i=1}^{n} d_i \left\| W^T x_i + b - y_i - y_i \circ m_i \right\|_2^2 + \gamma \left\| W \right\|_F^2 \tag{24}$$

where

$$d_i = \begin{cases} \frac{p}{2} g_i(W,b,m_i)^{\frac{p-2}{2}}, & g_i(W,b,m_i)^{\frac{p}{2}} \leq \varepsilon \\ 0, & otherwise \end{cases} \tag{25}$$

according to the first step of Algorithm 1 and Eq. (23).

When the M is fixed, the problem (24) can be written in matrix form as:

$$\begin{aligned} \min_{W,b} \quad & Tr((W^T X + b\mathbf{1}^T - Z) D (W^T X + b\mathbf{1}^T - Z)^T) \\ & + \gamma \left\| W \right\|_F^2 , \end{aligned} \tag{26}$$

where $\mathbf{1} \in \Re^{n\times 1}$ is the vector with all elements as 1, $Z \in \Re^{c\times n}$ with the i-th column as $z_i = y_i + y_i \circ m_i$, and D is a diagonal matrix with the i-th diagonal element as d_i defined in Eq. (25).

By setting the derivative of Eq. (26) w.r.t. b to zeros, we have:

$$b = \frac{1}{\mathbf{1}^T D \mathbf{1}} Z D \mathbf{1} - \frac{1}{\mathbf{1}^T D \mathbf{1}} W^T X D \mathbf{1} . \tag{27}$$

Substituting Eq. (27) into Eq. (26), the problem (26) becomes:

$$\min_{W} Tr(W^T (XHX^T + \gamma I) W) - 2Tr(W^T X H Z^T) , \tag{28}$$

where $I \in \Re^{d\times d}$ is an identity matrix.

By setting the derivative of Eq. (28) w.r.t. W to zeros, we have:

$$W = (XHX^T + \gamma I)^{-1} X H Z^T , \tag{29}$$

where $H = D - \frac{1}{\mathbf{1}^T D \mathbf{1}} D \mathbf{1}\mathbf{1}^T D$.

When the W and b are fixed, the problem (24) can be solved by separately solving the following problem for each m_i:

$$\min_{m_i \geq 0} \left\| W^T x_i + b - y_i - y_i \circ m_i \right\|_2^2 . \tag{30}$$

Note that $y_i \in \pm 1$, the problem (30) can be equivalently rewritten as:

$$\min_{m_i \geq 0} \left\| y_i \circ (W^T x_i) + y_i \circ b - \mathbf{1} - m_i \right\|_2^2 , \tag{31}$$

where $\mathbf{1} \in \Re^{c\times 1}$ is the vector with all elements as 1. It can be easily seen that the optimal solution to the problem (31) is:

$$m_i = (y_i \circ (W^T x_i) + y_i \circ b - \mathbf{1})_+ , \tag{32}$$

where the k-th element of vector $(v)_+$ is $max(v_k, 0)$.

Based on the above analysis, the detailed algorithm to solve the multiclass capped ℓ_p-norm SVM problem (12) is summarized in Algorithm 2.

Algorithm 2 Algorithm to solve the problem (12).

Initialize $D = I$. Initialize all the elements of M as 0.

while not converge **do**

1. Calculate Z, where the i-th column $z_i = y_i + y_i \circ m_i$
2. Update W by Eq. (29):
$W = (XHX^T + \gamma I)^{-1} X H Z^T$
3. Update b by Eq. (27):
$b = \frac{1}{\mathbf{1}^T D \mathbf{1}} Z D \mathbf{1} - \frac{1}{\mathbf{1}^T D \mathbf{1}} W^T X D \mathbf{1}$
4. Update M, where the i-th column m_i is by Eq. (32):
$m_i = (y_i \circ (W^T x_i) + y_i \circ b - \mathbf{1})_+$
5. Update the diagonal matrix D, where the i-th diagonal element d_i is calculated by Eq. (25):

$$d_i = \begin{cases} \frac{p}{2} g_i(W,b,m_i)^{\frac{p-2}{2}}, & g_i(W,b,m_i)^{\frac{p}{2}} \leq \varepsilon \\ 0, & otherwise \end{cases}$$

end while

Experimental Results

To experimentally validate the classification ability of our proposed method, in this section we will compare with five related methods and exhibit the experimental results on six benchmark datasets. For simplicity, we denote our capped ℓ_p SVM method as CappedSVM in the following context.

Datasets Description

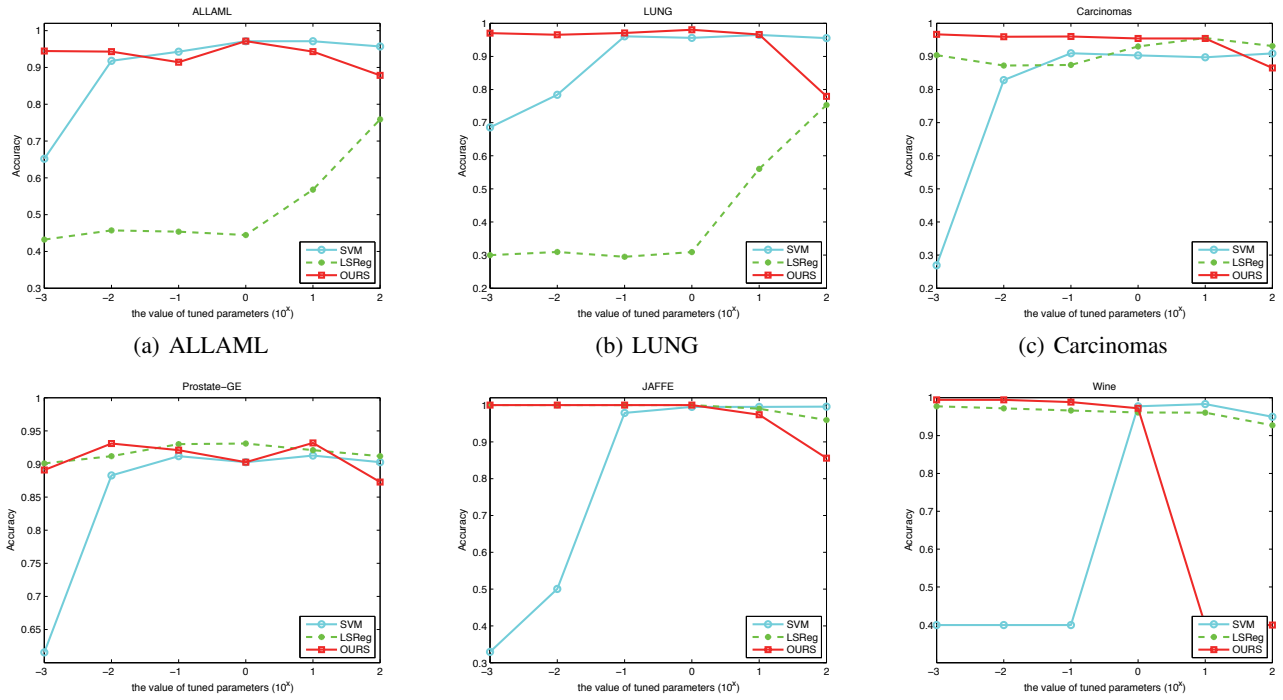
The six benchmark datasets involved in our experiments are: ALLAML data set (Fodor 1997), the Human Lung Carcinomas (LUNG) data set (Bhattacharjee et al. 2001), the Human Carcinomas (Carcinomas) data set (Su et al. 2001), the Prostate Cancer Gene Expression (Prostate-GE) data set (Singh et al. 2002), the Japanese Female Facial Expression (JAFFE) data set (Lyons, Kamachi, and Gyoba 1997), the chemical analysis of wine (Wine) data set, and the physical measurements of abalone (Abalone) data set, the first four of which are gene expression microarray data sets, the latter one is an image data set while the last one is a daily life data set from the UCI Machine Learning Repository (Bache and Lichman 2013). The property of these six data sets is summarized in Table 1.

Table 1: Description of Datasets

Datasets	# of Samples	Features	Classes
ALLAML	72	7129	2
LUNG	203	3312	5
Carcinomas	174	9182	11
Prostate-GE	102	5966	2
JAFFE	213	1024	10
Wine	178	13	3

Classification Results Comparison

We compared our classification method with k-Nearest Neighbors algorithm (KNN), Support Vector Machine (SVM) and the Least Square Support Vector Machine



(a) ALLAML (b) LUNG (c) Carcinomas

(d) Prostate (e) JAFFE (f) Wine

Figure 2: Classification Accuracy under Different Parameter Settings

(LSSVM) (Suykens et al. 2002). We exploited the program from LIBSVM[2] for the linear SVM method.

Before classification, all data sets are normalized to the range of [0, 1] and randomly divided using 5-fold cross validation.

The evaluation of different methods is based on the classification accuracy. For the KNN method, we set $k = 1$. For all other methods involving a parameter, including SVM, LSSVM and CappedSVM, we tuned the parameter to be $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ separately and recorded the best results. In our method, we set the value of ε in a heuristic way, that is, in the first five iterations, we selected 10% data with the largest noise to determine ε.

From Table 2, we can come into the conclusion that our proposed method works very well on real benchmark data sets. Our Capped SVM method has a high potential to outperform other traditional methods on these various kinds of data sets. In addition, we compared the classification accuracy of different methods under different parameters. The comparison results in Fig. 2 confirms the stability of our method. Compared with SVM and LSSVM, our method is more robust to the setting of parameters, which alleviates the burden of tuning parameters to a large extent.

Datasets	KNN	SVM	LSReg	CappedSVM
ALLAML	87.32	97.13	75.89	**97.14**
LUNG	94.64	96.47	75.95	**98.03**
Carcinomas	82.51	90.96	95.54	**96.66**
Prostate-GE	79.36	91.27	93.09	**93.18**
JAFFE	99.51	99.59	**100.00**	**100.00**
Wine	95.49	98.33	97.75	**99.44**

Table 2: Classification Accuracy (%) on Real Benchmark Datasets

Conclusions

Although the SVM models have been successfully applied to solve numerous classification tasks, the hinge loss and squared hinge loss used in existing SVM methods are not robust to data outliers, which often exist in the real-world applications. To tackle this problem, we proposed a novel capped ℓ_p SVM classification model by utilizing a new capped ℓ_p-norm based objective. Our capped ℓ_p-norm based objective is theoretically and empirically robust to data outliers. To solve the new objective, we derived new efficient optimization algorithms with rigorously proved convergence. The experimental results on six benchmark datasets show that our new capped ℓ_p SVM method is robust to data outliers and consistently improve the classification performance.

[2]http://www.csie.ntu.edu.tw/˜cjlin/libsvm/.

References

Bache, K., and Lichman, M. 2013. UCI machine learning repository.

Bhattacharjee, A.; Richards, W. G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.;

Gillette, M.; et al. 2001. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* 98(24):13790–13795.

Boser, B. E.; Guyon, I.; and Vapnik, V. 1992. A training algorithm for optimal margin classifiers. *Proc. Fifth ACM Workshop on Computational Learning Theory (COLT)* 144–152.

Bradley, P., and Mangasarian, O. 1998. Feature selection via concave minimization and support vector machines. *International Conference on Machine Learning (ICML)*.

Brown, M. P. S.; Grundy, W. N.; Lin, D.; et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* 97:262–267.

Cai, X.; Nie, F.; Huang, H.; and Ding, C. H. Q. 2011. Multi-class l2,1-norm support vector machine. In *IEEE 11th International Conference on Data Mining (ICDM)*, 91–100.

Cao, H.; Hu, D. H.; Shen, D.; Jiang, D.; Sun, J.-T.; Chen, E.; and Yang, Q. 2009. Context-aware query classification. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, 3–10.

Chang, K.-W.; Hsieh, C.-J.; and Lin, C.-J. 2008. Coordinate descent method for large-scale L2-loss linear svm. *Journal of Machine Learning Research* 9:1369–1398.

Chen, C.; Daponte, J.; and Fox, M. 1989. Fractal feature analysis and classification in medical imaging. *IEEE Trans Med Imaging* 8(2):133–42.

Cotter, A.; Shalev-Shwartz, S.; and Srebro, N. 2013. Learning optimally sparse support vector machines. *International Conference on Machine Learning (ICML)* 266–274.

Fodor, S. P. 1997. Dna sequencing: Massively parallel genomics. *Science*.

Gao, H.; Nie, F.; Cai, W.; and Huang, H. 2015. Robust capped norm nonnegative matrix factorization. *24th ACM International Conference on Information and Knowledge Management (CIKM 2015)* 871–880.

Gong, P.; Ye, J.; and Zhang, C. 2013. Multi-stage multi-task feature learning. *J. Mach. Learn. Res.* 14(1):2979–3010.

Hsieh, C.-J.; Chang, K.-W.; Keerthi, S. S.; Sundararajan, S.; and Lin, C.-J. 2008. A dual coordinate descent method for large-scale linear svm. In *International Conference on Machine Learning (ICML)*, 408–415.

Huo, Z.; Nie, F.; and Huang, H. 2016. Robust and effective metric learning using capped trace norm. *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016)* 1605–1614.

Jiang, W.; Nie, F.; and Huang, H. 2015. Robust dictionary learning with capped l1 norm. *Twenty-Fourth International Joint Conferences on Artificial Intelligence (IJCAI 2015)* 3590–3596.

Keerthi, S. S., and DeCoste, D. 2005. A modified finite newton method for fast solution of large scale linear svms. *Journal of Machine Learning Research* 6:341–361.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems 25*. 1097–1105.

Lin, C.-J.; Weng, R. C.; and Keerthi, S. S. 2008. Trust region newton method for large-scale logistic regression. *Journal of Machine Learning Research* 9:627–650.

Lyons, M.; Kamachi, M.; and Gyoba, J. 1997. Japanese female facial expressions (jaffe). *Database of digital images*.

Mangasarian, O. L. 2002. A finite newton method for classification. *Optimization Methods and Software* 17(5):913–929.

Singh, D.; Febbo, P. G.; Ross, K.; Jackson, D. G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A. A.; D'Amico, A. V.; Richie, J. P.; et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell* 1(2):203–209.

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanogl, H.; and Demirbas, M. 2010. Short text classification in twitter to improve information filtering. *the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR)* 841–842.

Su, A. I.; Welsh, J. B.; Sapinoso, L. M.; Kern, S. G.; Dimitrov, P.; Lapp, H.; Schultz, P. G.; Powell, S. M.; Moskaluk, C. A.; Frierson, H. F.; et al. 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research* 61(20):7388–7393.

Sun, Q.; Xiang, S.; and Ye, J. 2013. Robust principal component analysis via capped norms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, 311–319.

Suykens, J. A.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J.; Suykens, J.; and Van Gestel, T. 2002. *Least squares support vector machines*, volume 4. World Scientific.

Wang, S., and Mannin, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. *the 50th Annual Meeting of the Association for Computational Linguistics (ACL)* 90–94.

Wang, L.; Zhu, J.; and Zou, H. 2007. Hybrid huberized support vector machines for microarray classification. *International Conference on Machine Learning (ICML)*.

Xiang, S.; Nie, F.; Meng, G.; Pan, C.; and Zhang, C. 2012. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learning Syst.* 23(11):1738–1754.

Zhang, T. 2008. Multi-stage convex relaxation for learning with sparse regularization. *Advances in Neural Information Processing Systems*.

Zhang, T. 2010. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* 1081–1107.