Multi-Class Learning using Unlabeled Samples: Theory and Algorithm

Jian Li, Yong Liu*, Rong Yin, and Weiping Wang

{lijian9026, liuyong, yinrong, wangweiping}@iie.ac.cn

中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Introduction

In this paper, we investigate the generalization performance of **multi-class classification (MC)**, for which we obtain a shaper error bound by using the notion of **local Rademacher complexity (LRC)** and **additional unlabeled samples (SSL)**, substantially improving the state-of-the-art bounds in existing multi-class learning methods. The statistical learning motivates us to devise an efficient multi-class learning framework with the local Rademacher complexity and Laplacian regularization. Coinciding with the theoretical analysis, experimental results demonstrate that the stated approach achieves better performance.

1. Core Idea: Linear-MC + LRC + SSL.
2. Theory
(1) Label-dependent complexity $\Rightarrow$ Label-independent complexity (use both label and unlabeled samples) : $\mathcal{R}_n(\mathcal{L}_r) \Rightarrow \mathcal{R}(\mathcal{H}_r)$
(2) Shaper generalization error bounds with convergence rate: $\mathcal{O}(\frac{K}{\sqrt{n+u}}+\frac{1}{n})$ or $\mathcal{O}(\frac{1}{n})$.
3. Algorithm
(1) Multi-penalty minimization

$$\arg\min_{h\in\mathcal{H}_r} \frac{1}{n}\sum_{i=1}^{n}\ell(h(\mathbf{x}_i),y_i)+\tau_A\|\mathbf{W}\|_F^2 + \tau_I \mathrm{trace}(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W})+\tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})$$

A. $\mathrm{trace}(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W})$: Laplacian regularization to make use of unlabeled examples.
B. $\sum_{j>\theta}\lambda_j(\mathbf{W})$: the tail sum of singular values to bound LRC.
(2) Optimization algorithm: SGD and partly singular values thresholding (SVT).

Problem Definition

To evaluate the label of $\mathbf{x}$, we wish to learn a scoring rule from the hypothesis space $\mathcal{H}$

$$h(\mathbf{x})=\mathbf{W}^T\mathbf{x},$$

where $h\in\mathcal{H}, \mathbf{W}\in\mathbb{R}^{d\times K}$ and $\mathbf{x}\in\mathbb{R}^d$, thus h is a vector-valued function with mapping $\mathcal{X}\to\mathbb{R}^K$. The predictor uses the following mapping to predict labels $\mathbf{x}\to\arg\max_y h(\mathbf{x},y)$, where $h(\mathbf{x},y)=[\mathbf{W}^T\mathbf{x}]_y$ means the y-th value in vector $\mathbf{W}^T\mathbf{x}$. For any hypothesis $h\in\mathcal{H}$, the margin of a labeled example $(\mathbf{x},y)$ is defined as

$$\rho_h(\mathbf{x},y)=h(\mathbf{x},y)-\max_{y\neq y'}h(\mathbf{x},y').$$

The loss space associated with $\mathcal{H}$ is defined as $\mathcal{L}=\{\ell(\rho_h(\mathbf{x},y))|h\in\mathcal{H}\}$. Under assumptions:
(1) The loss function is continuous and bounded.
(2) ℓ is L-Lipschitz continuous w.r.t. $\rho_h(\mathbf{x},y)$.

Definition 1. *The empirical Rademacher complexity of loss space and hypotheses space*

$$\widehat{\mathcal{R}}_n(\mathcal{L}_r)=\mathbb{E}_\sigma\sup_{\ell\in\mathcal{L}_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(h(\mathbf{x}_i,y_i)),$$

$$\widehat{\mathcal{R}}(\mathcal{H}_r)=\mathbb{E}_\sigma\sup_{h\in\mathcal{H}_r}\frac{1}{n+u}\sum_{i=1}^{n+u}\sigma_i h(\mathbf{x}_i,y_i^\circ),$$

where $\sigma_1,\sigma_2,\cdots,\sigma_{n+u}$ are $\{\pm1\}$-valued independent Rademacher random variables.

Theory

Theorem 1. *For any $\ell\in\mathcal{L}_r:\mathcal{X}\times\mathcal{Y}\to[0,1]$, consider a sub-root function $\psi(r)$ with fixed point r^* and such that $\forall r>r^*$, $KL\mathcal{R}(\mathcal{H}_r)\leq\psi(r)$, then $\forall\ell\in\mathcal{L}_r$ and $\forall k>1$, with probability at least $1-\delta$*

$$L(\ell)\leq\max\left\{\frac{k}{k-1}\widehat{L}(\ell),\widehat{L}(\ell)+c_4r^*+\frac{c_1}{n}\right\},$$

where $c_1=(3+4k)\log(1/\delta), c_4=32k$.

Theorem 2. *Let $\mathbf{W}=\mathbf{U\Sigma V}$ be SVD decomposition of $\mathbf{W}$,$\mathbf{U}$ and $\mathbf{V}$ are unitary matrices with size of $d\times d$ and $K\times K$ respectively, and $\mathbf{\Sigma}$ is a $d\times K$ matrix with singular values $\{\lambda_j\}$ on the diagonal in descending order.*

$$\mathcal{R}(\mathcal{H}_r)\leq\frac{1}{KL}\sqrt{\frac{r\theta}{n+u}}+\frac{\sum_{j>\theta}\lambda_j}{\sqrt{n+u}}.$$

Theorem 3. *For any $\ell\in\mathcal{L}_r:\mathcal{X}\times\mathcal{Y}\to[0,1]$, $\forall k>1,\|\mathbf{W}\|\leq1$ and $\forall\ \delta\in(0,1)$, the following holds with probability at least $1-\delta$,*

$$L(\ell)\leq\max\left\{\frac{k}{k-1}\widehat{L}(\ell),\ \widehat{L}(\ell)+\frac{c_1}{n}+\frac{c_2}{n+u}+\frac{c_3K\sum_{j>\theta}\lambda_j(\mathbf{W})}{\sqrt{n+u}}\right\},$$

where $c_1=(3+4k)\log(1/\delta)$, $c_2=32k\theta$ and $c_3=64kL$, $\lambda_j(\mathbf{W})$ is the j largest singular value of matrix $\mathbf{W}$.

Bounds	Common Case	Special Case
[Allwein et al. 2000]	$\mathcal{O}(\frac{\sqrt{V}\log K}{\sqrt{n}})$	
[Cortes et al. 2013]	$\mathcal{O}(\frac{K}{\sqrt{n}})$	
[Maximov et al. 2018]†	$\mathcal{O}(\sqrt{\frac{K}{n}}+K\sqrt{\frac{K}{u}})$	
[Li et al. 2018]	$\mathcal{O}((c_1+c_2)\frac{\log^2K}{n})$	
Theorem 3†	$\mathcal{O}(\frac{K}{\sqrt{n+u}}+\frac{1}{n})$	$\mathcal{O}(\frac{c_1}{n})$

Table 1: Comparison of multi-class classification error bounds, including one VC-dimension bound, two global Rademacher complexity bounds, and two local Rademacher complexity bounds. Here $n\ll u, K\ll n$ and † represents using unlabeled data.

Algorithm

For the sake of simplification, we rewrite the optimization as

$$\arg\min_{h\in\mathcal{H}_r}\ \tau_S\sum_{j>\theta}\lambda_j(\mathbf{W})+g(\mathbf{W})\qquad\text{where}$$

$$g(\mathbf{W})=\frac{1}{n}\sum_{i=1}^{n}\overbrace{|1-([\mathbf{W}^T\mathbf{x}_i]_{y_i}-\max_{y'\neq y_i}[\mathbf{W}^T\mathbf{x}_i]_{y'})|_+}^{\omega(\mathbf{W},\mathbf{x}_i)} +\tau_A\|\mathbf{W}\|_F^2+\tau_I\mathrm{trace}(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W}).$$

1. Stochastic Gradient Descent (SGD)

$$\nabla\omega(\mathbf{W},\mathbf{x}_i)=\begin{cases}\mathbf{0}, & [\mathbf{W}^T\mathbf{x}_i]_{y_i}-\max_{y'\neq y_i}[\mathbf{W}^T\mathbf{x}_i]_{y'}\geq1,\\ [0,\cdots,\underbrace{-\mathbf{x}_i}_{y_i},\cdots,\underbrace{\mathbf{x}_i}_{y'},\cdots,0]_{d\times K}, & \text{else.}\end{cases}$$

GD update gradients on the entire dataset

$$\nabla g(\mathbf{W})=\frac{1}{n}\sum_{i=1}^{n}\nabla\omega(\mathbf{W},\mathbf{x}_i)+2\tau_A\mathbf{W}+2\tau_I\mathbf{XLX}^T\mathbf{W}.$$

SGD update gradients on a random sample $\mathbf{x}'$

$$\nabla g(\mathbf{W},\mathbf{x}')=\nabla\omega(\mathbf{W},\mathbf{x}')+2\tau_A\mathbf{W}+2\tau_I\mathbf{XLX}^T\mathbf{W}.$$

2. Partly Singular Value Thresholding
Compute SVD decomposition

$$\mathbf{U\Sigma V}^T=\mathbf{W}^t-\frac{1}{\mu}\nabla g(\mathbf{W}^t,\mathbf{x}_{i_t})$$

Update $\mathbf{W}^{t+1}$ using Proposition 1

$$\mathbf{W}^{t+1}=\mathbf{U\Sigma}^\theta_{\frac{\tau_S}{\mu}}\mathbf{V}^T.$$

Proposition 1 (Theorem 6 of [Xu et al. 2016]). *Let $\mathbf{Q}\in\mathbb{R}^{d\times K}$ with rank r and its SVD decomposition is $\mathbf{Q}=\mathbf{U\Sigma V}^T$, where $\mathbf{U}\in\mathbb{R}^{d\times r}$ and $\mathbf{V}\in\mathbb{R}^{K\times r}$ are orthogonal, $\mathbf{\Sigma}$ is diagonal. Then,*

$$\mathcal{D}^\theta_\tau(\mathbf{Q})=\arg\min_{\mathbf{W}}\left\{\frac{1}{2}\|\mathbf{W}-\mathbf{Q}\|_F^2+\tau\sum_{j>\theta}\lambda_j(\mathbf{W})\right\},$$

is given by $\mathcal{D}^\theta_\tau=\mathbf{U\Sigma}^\theta_\tau\mathbf{V}^T$, where $\mathbf{\Sigma}^\theta_\tau$ is diagonal

$$(\mathbf{\Sigma}^\theta_\tau)_{jj}=\begin{cases}\max(0,\Sigma_{jj}-\tau), & i\leq\theta,\\ \mathbf{\Sigma}_{jj}, & i>\theta.\end{cases}$$

Experimental Results

Approaches	Linear-MC	LRC-MC	SS-MC	PS3VT
iris	27.12±5.36	24.57±6.13	**23.53±5.04**	23.71±5.22
wine	8.77±3.22	8.33±5.22	8.20±4.12	**7.63±3.88**
glass	48.68±5.32	47.46±5.40	46.68±4.83	**46.28±5.18**
svmguide2	23.31±3.86	22.42±3.68	22.33±3.99	**21.37±3.46**
vowel	47.40±3.73	47.05±2.89	46.66±3.36	**45.74±3.15**
vehicle	33.78±2.17	29.74±2.41	29.67±2.73	**28.53±2.48**
dna	8.83±0.94	8.69±0.86	8.56±0.78	**8.56±0.78**
segment	26.69±2.20	26.84±2.37	26.28±2.30	**26.09±2.20**
satimage	15.94±0.83	**15.88±0.83**	15.92±0.87	15.89±0.82
pendigits	10.22±0.89	8.37±0.53	7.24±0.44	**6.46±0.37**
usps	7.19±0.42	7.09±0.41	7.10±0.41	**7.06±0.45**
shuttle	23.25±0.32	21.61±0.31	21.55±0.28	**21.48±0.28**
letter	28.31±0.54	26.98±0.49	26.92±0.52	**26.91±0.48**
poker	52.34±0.50	50.30±0.38	50.22±0.40	**50.11±0.45**
Sensorless	54.71±1.26	54.04±1.46	53.15±1.43	**52.50±1.22**

Table 2: Comparison of test err (%) among our proposed `PS3VT` and other linear Multi-class classification methods. For each dataset, we bold the optimal test error and underline results in other methods which show no significant difference from the optimal one.