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

Multi-Class Learning using Unlabeled Samples: Theory and Algorithm

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

Institute of Information Engineering, Chinese Academy of Sciences

28th International Joint Conference on Artificial Intelligence (IJCAI 2019)

Contents

Brief

1. Core Idea: Linear-MC + LRC + SSL
   - Linear-MC: Linear max-margin multi-class classification estimator
   - LRC: Local Rademacher complexity
   - SSL: Semi-supervised learning (additional unlabeled samples)
2. Theory: Shaper generalization error bounds with convergence rate
   - In the worst case: $\mathcal{O}\left(\frac{K}{\sqrt{n+u}}+\frac{1}{n}\right)$
   - In the benign cases: $\mathcal{O}\left(\frac{1}{n}\right)$

   Rate of common GRC bounds: $\mathcal{O}(\frac{K}{\sqrt{n}})$
3. Algorithm: An unified learning framework
   - Multi-penalty Objective

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

   - Optimization algorithm: SGD and partly singular values thresholding

Linear max-margin multi-class classification

1. Semi-supervised multi-class classification data setting
   - n: the number of labeled examples.
   - u: the number of unlabeled examples. Typically, $n \ll u$.
   - K: the number of categories. Output space $\mathcal{Y} = \{1, 2, \cdots, K\}$.
   - d: the feature dimension of input sample. Input space $\mathcal{X} = \mathbb{R}^d$.
2. The estimator (to learn)

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

   where $\mathbf{W} \in \mathbb{R}^{d \times K}, \mathbf{x} \in \mathbb{R}^d$ and $h : \mathcal{X} \to \mathbb{R}^K$.
3. The predictor

$$\mathbf{x} \to \arg\max_y h(\mathbf{x}, y),$$

   where $h(\mathbf{x}, y) = [\mathbf{W}^T \mathbf{x}]_y$ means the y-th value in vector $\mathbf{W}^T \mathbf{x}$.
4. The margin

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

Loss space and generalization errors

1. Loss function: the estimator h misclassifies example $(\mathbf{x}, y)$ if $\rho_h(\mathbf{x}, y) \leq 0$.
   - 0-1 loss: $\ell(\rho_h(\mathbf{x}, y)) = 1_{\rho_h(\mathbf{x},y)\leq 0}$.
   - Hinge loss: $\ell(\rho_h(\mathbf{x}, y)) = |1 - \rho_h(\mathbf{x}, y)|_+$.
   - Squared hinge loss $\ell(\rho_h(\mathbf{x}, y)) = (1 - \rho_h(\mathbf{x}, y))_+^2$.
2. Standard assumption:
   the loss function ℓ is L-Lipschitz continuous w.r.t. $\rho_h(\mathbf{x}, y)$.
3. Expected loss and empirical loss

$$L(\ell) = \mathbb{E}_\mu[\ell(\rho_h(\mathbf{x}, y))], \quad \widehat{L}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \ell(\rho_h(\mathbf{x}, y)).$$

4. Loss space: given by ℓ and hypotheses space $\mathcal{H}$

$$\mathcal{L} = \{\ell(\rho_h(\mathbf{x}, y)) | h \in \mathcal{H}\}.$$

5. Generalization errors bound (Uniform Deviation)
   difference between the expected loss and the empirical loss

$$\sup_{\ell \in \mathcal{L}} L(\ell) - \widehat{L}(\ell)$$

Global Rademacher complexity

1. In learning theory, *global* Rademacher complexity is one of the most successful **data-dependent** tools to measure the richness of the classes of all functions.

Definition

The empirical *global* Rademacher complexity of the class of functions F is

$$\widehat{\mathcal{R}}_n(F) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i),$$

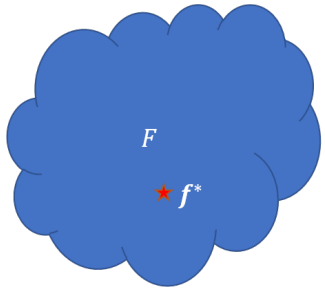
where $\sigma_1, \cdots, \sigma_n$ are independent Rademacher variables, for which $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. The expected version is $\mathcal{R}_n(F) = \mathbb{E}\, \widehat{\mathcal{R}}_n(F)$.

2. The generalization error bound can be controlled by the Rademacher averages of loss space [Bartlett et al., 2005].

$$\mathbb{E}\left[\sup_{\ell \in \mathcal{L}} L(\ell) - \widehat{L}(\ell)\right] \leq 2\mathbb{E}\, \mathcal{R}_n(\mathcal{L}).$$

Local Rademacher complexity

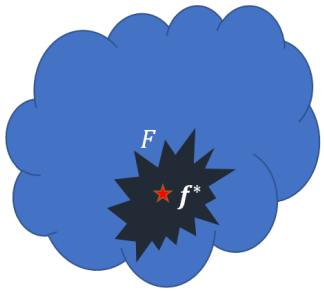
1. *Global* Rademacher complexity (GRC): measures the complexity on the set of all functions.



Global Rademacher complexity:
on entire function class F

2. *Local* Rademacher complexity (LRC): measures the complexity on a favorable subset of function classes.

$\mathcal{R}_n(F_r)$ with a smaller subset $F_r = \{f \in F : \mathrm{Var}(f) \leq r\}$



Local Rademacher complexity:
on a small subset of F

Local Rademacher complexity usually leads to a **better performance**.

Contents

How to make use of unlabeled examples?

1. Label-dependent $\mathcal{R}_n(\mathcal{L}_r)$

$$\begin{aligned}\mathcal{R}_n(\mathcal{L}_r) &= \mathcal{R}_n(\{\ell \in \mathcal{L} : \mathsf{Var}(\ell) \in r\}) \\ &= \mathbb{E} \sup_{\ell \in \mathcal{L}:\mathsf{Var}(\ell)\in r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(\rho_h(x_i, y_i)).\end{aligned}$$

The estimate of $\mathcal{R}_n(\mathcal{L}_r)$ is related to n, of which rate is $\mathcal{O}(\frac{1}{n})$ or $\mathcal{O}(\frac{1}{\sqrt{n}})$.

2. Label-independent $\mathcal{R}(\mathcal{H}_r)$

$$\mathcal{R}(\mathcal{H}_r) = \mathbb{E} \sup_{h \in \mathcal{H}_r} \frac{1}{n+u} \sum_{i=1}^{n+u} \sigma_i h(x_i).$$

The estimate of $\mathcal{R}_n(\mathcal{L}_r)$ is related to $n+u$, of which rate is $\mathcal{O}(\frac{1}{n+u})$ or $\mathcal{O}(\frac{1}{\sqrt{n+u}})$.

General semi-supervised LRC bound

1. Connection between $\mathcal{R}_n(\mathcal{L}_r)$ and $\mathcal{R}(\mathcal{H}_r)$

$$\mathcal{R}_n(\mathcal{L}_r) \leq KL\mathcal{R}(\mathcal{H}_r)$$

   due to L-Lipschitz continuous condition.
2. General generalization error bound (Extension of [Bartlett et al., 2005])

Theorem

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* $1 - \delta$

$$L(\ell) \leq \max\left\{\frac{k}{k-1}\widehat{L}(\ell), \widehat{L}(\ell) + c_4 r^* + \frac{c_1}{n}\right\},$$

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

Estimate $\mathcal{R}(\mathcal{H}_r)$ with SVD

Theorem

Let $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$ 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. Assume $\|\mathbf{W}\| \leq 1$, such that the local Rademacher complexity $\mathcal{R}(\mathcal{H}_r)$ over all examples is upper bounded by

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

1. Based on Singular Value Decomposition (SVD) on the weighted matrix $\mathbf{W}$ [Yu et al., 2014, Xu et al., 2016].
2. LRC is determined by tail sum of singular values of $\mathbf{W}$, that is $\sum_{j>\theta}\lambda_j$.
3. When $\theta = 0$, LRC degrades into GRC , that is the trace of $\mathbf{W}$.

Explicit LRC error bound for multi-class classification

Theorem

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

$$L(\ell) \leq \max\left\{\frac{k}{k-1}\widehat{L}(\ell), \widehat{L}(\ell) + \frac{c_1}{n} + \frac{c_2}{n+u} + \frac{c_3 K \sum_{j>\theta}\lambda_j}{\sqrt{n+u}}\right\},$$

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

1. The convergence rate of the bound is no faster than $\mathcal{O}(\frac{1}{n})$.
2. The convergence rate is determined by the second term and fourth term.
   - In the worst case: $\mathcal{O}\left(\frac{K}{\sqrt{n+u}} + \frac{1}{n}\right)$.
     e.g. GRC when $\theta = 0$.
   - In the benign cases: $\mathcal{O}\left(\frac{1}{n}\right)$.
     e.g. Singular values of $\mathbf{W}$ decay exponentially quickly.

Comparison with other multi-class bounds

Bounds	Common Case	Special Case
VC-deminsion [Allwein et al., 2000]	$\mathcal{O}\left(\frac{\sqrt{V}\log K}{\sqrt{n}}\right)$	
GRC [Cortes et al., 2013]	$\mathcal{O}\left(\frac{K}{\sqrt{n}}\right)$	
GRC [Maximov et al., 2018]†	$\mathcal{O}\left(\sqrt{\frac{K}{n}}+K\sqrt{\frac{K}{u}}\right)$	
LRC for kernel-MC [Li et al., 2018]	$\mathcal{O}\left((c_1+c_2)\frac{\log^2 K}{n}\right)$	
Theorem 4†	$\mathcal{O}\left(\frac{K}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{c_1}{n}\right)$

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 making use of unlabeled data.

Contents

Multi-penalty objective

- Multi-penalty objective

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

  where loss function is $\ell(h(\mathbf{x}_i),y_i)=|1-(h(\mathbf{x}_i,y_i)-\max_{y'\neq y_i}h(\mathbf{x}_i,y'))|_+$, $\lambda_j(\mathbf{W})$ denotes the j-th largest singular value of $\mathbf{W}\in\mathbb{R}^{d\times K}$.

- For the sake of simplification, we rewrite the optimization as

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

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

1. Stochastic Gradient Descent (SGD)

1. Sub-gradient on empirical risk

$$\nabla\omega(\mathbf{W},\mathbf{x}_i)=\begin{cases}\mathbf{0}, & [\mathbf{W}^T\mathbf{x}_i]_{y_i}-\max\limits_{y'\neq y_i}[\mathbf{W}^T\mathbf{x}_i]_{y'}\geq 1,\\ [0,\cdots,\underbrace{-\mathbf{x}_i}_{y_i},\cdots,\underbrace{\mathbf{x}_i}_{y'},\cdots,0]_{d\times K}, & \text{else.}\end{cases}$$

   - GD update gradients on the entire dataset

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

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

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

2. Update $\mathbf{W}$

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

3. Compute SVD decomposition

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T=\mathbf{W}^t.$$

2. Partly Singular Value Thresholding

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

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

Proposition (Theorem 6 of [Xu et al., 2016])

Let $\mathbf{Q} \in \mathbb{R}^{d\times K}$ with rank r and its SVD decomposition is $\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d\times r}$ and $\mathbf{V} \in \mathbb{R}^{K\times r}$ are orthogonal, $\boldsymbol{\Sigma}$ 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 $\mathcal{D}^{\theta}_{\tau} = \mathbf{U}\boldsymbol{\Sigma}^{\theta}_{\tau}\mathbf{V}^T$, where $\boldsymbol{\Sigma}^{\theta}_{\tau}$ is diagonal

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

Contents

Compared methods

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

Parameters	Optimization objectives
$\tau_I=0,\tau_S=0$	Linear-MC [Koltchinskii et al., 2001]
$\tau_I=0,\tau_S>0$	LRC-MC [Li et al., 2018]
$\tau_I>0,\tau_S=0$	SS-MC [Li et al., 2015]
$\tau_I>0,\tau_S>0$	PS3VT

Table 2: Connections with other algorithms

Comparison of Test Error

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

Contents

Conclusion

1. Core Idea: Linear-MC + LRC + SSL
   - Linear-MC: Linear max-margin multi-class classification estimator
   - LRC: Local Rademacher complexity
   - SSL: Semi-supervised learning (additional unlabeled samples)
2. Theory: Shaper generalization error bounds with convergence rate
   - In the worst case: $\mathcal{O}\left(\frac{K}{\sqrt{n+u}}+\frac{1}{n}\right)$
   - In the benign cases: $\mathcal{O}\left(\frac{1}{n}\right)$
3. Algorithm: Unified learning framework
   - Multi-penalty Objective: empirical risk + model complexity

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

   - Optimization algorithm: SGD and partly singular values thresholding

References

Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 46–54, 2013.

Vladimir Koltchinskii, Dmitriy Panchenko, and Fernando Lozano. Some new bounds on the generalization error of combined classifiers. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 245–251, 2001.

Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems 31 (NIPS)*, pages 1591–1600, 2018.

Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4211–4219, 2015.

Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61:761–786, 2018.

Chang Xu, Tongliang Liu, Dacheng Tao, and Chao Xu. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 593–601, 2014.

Any Questions?