Multi-Class Learning: From Theory to Algorithm

Jian Li[1,2], **Yong Liu**[1,*], Rong Yin[1,2], Hua Zhang[1], Lizhong Ding[5], Weiping Wang[1,3,4]

[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]National Engineering Research Center for Information Security
[4]National Engineering Laboratory for Information Security Technology
[5]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE

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

Introduction

- Statistical learning of multi-class classification is a crucial problem in machine learning.
- Existing generalization bounds for multi-class classification:

Methods	Convergence rate
VC-dimension	$\mathcal{O}(\sqrt{\mathbf{V}}\log \mathbf{K}/\sqrt{\mathbf{n}})$
Natarajan dimension	$\mathcal{O}(\mathbf{d}_{\mathrm{Nat}}/\mathbf{n})$
Covering Number	$\mathcal{O}(1/\sqrt{\mathbf{n}})$
Rademacher Complexity	$\mathcal{O}(\log^2 \mathbf{K}/\sqrt{\mathbf{n}})$
Stability	$\mathcal{O}(1/\sqrt{\mathbf{n}})$
PAC-Bayesian	$\mathcal{O}(\sqrt{\hat{\mathbf{L}}(\mathbf{h}_\gamma)/\mathbf{n}})$

- Contributions:
  - A new local Rademacher complexity based bound with fast convergence rate $\mathcal{O}\big((\log \mathbf{K})^{2+1/\log \mathbf{K}}/\mathbf{n}\big)$ for multi-class classification is establish.
  - Two novel multi-class multiple kernel learning algorithms are proposed with statistical guarantees: a) `Conv-MKL`. b) `SMSD-MKL`.

Notations and Preliminaries

- Multi-class classification setting
  Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, 2, \dots, \mathbf{K}\}$ the output space. Based on training examples $\mathcal{S}$ drawn i.i.d. from a fixed, but unknown probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, we wish to learn a scoring rule $\mathbf{h}$ mapping from $\mathcal{Z}$ to $\mathbb{R}$ to predict $\mathbf{x} \to \arg\max_{\mathbf{y}\in\mathcal{Y}} \mathbf{h}(\mathbf{x},\mathbf{y})$. For any $\mathbf{h} \in \mathcal{H}$, the margin of a labeled example $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is defined as
$$\rho_{\mathbf{h}}(\mathbf{z}) := \mathbf{h}(\mathbf{x},\mathbf{y}) - \max_{\mathbf{y}'\neq\mathbf{y}} \mathbf{h}(\mathbf{x},\mathbf{y}').$$
  The $\mathbf{h}$ misclassifies the labeled example $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ if $\rho_{\mathbf{h}}(\mathbf{z}) \leq 0$. Let $\ell(\rho_{\mathbf{h}}(\mathbf{z}))$ be loss function, $\mathbf{L}(\ell_{\mathbf{h}})$ and $\hat{\mathbf{L}}(\ell_{\mathbf{h}})$ be expected generalization error and empirical error with respect to $\ell_{\mathbf{h}}$
$$\mathbf{L}(\ell_{\mathbf{h}}) := \mathbb{E}_\mu[\ell(\rho_{\mathbf{h}}(\mathbf{z}))] \text{ and } \hat{\mathbf{L}}(\ell_{\mathbf{h}}) = \frac{1}{\mathbf{n}}\sum_{\mathbf{i}=1}^{\mathbf{n}} \ell(\rho_{\mathbf{h}}(\mathbf{z}_{\mathbf{i}})).$$
- Hypothesis Space
  Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel with ϕ being the associated feature map. The $\ell_{\mathbf{p}}$-norm hypothesis space is denoted by:
$$\mathcal{H}_{\mathbf{p},\kappa} = \Big\{\mathbf{h}_{\mathbf{w}} = (\langle \mathbf{w}_1, \phi(\mathbf{x})\rangle, \dots, \langle \mathbf{w}_{\mathbf{K}}, \phi(\mathbf{x})\rangle) : \|\mathbf{w}\|_{2,\mathbf{p}} \leq 1, 1 \leq \mathbf{p} \leq 2\Big\},$$
  where $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_{\mathbf{K}})$ and $\|\mathbf{w}\|_{2,\mathbf{p}} = \Big[\sum_{\mathbf{i}=1}^{\mathbf{K}} \|\mathbf{w}_{\mathbf{i}}\|_2^{\mathbf{p}}\Big]^{1/\mathbf{p}}$ is the $\ell_{2,\mathbf{p}}$-norm. For $\mathbf{p} \geq 1$, the dual exponent $\mathbf{q}$ satisfies $1/\mathbf{p} + 1/\mathbf{q} = 1$. The space of loss function associated with $\mathcal{H}_{\mathbf{p},\kappa}$ is denoted by
$$\mathcal{L} = \{\ell_{\mathbf{h}} := \ell(\rho_{\mathbf{h}}(\mathbf{z})) : \mathbf{h} \in \mathcal{H}_{\mathbf{p},\kappa}\}.$$
- The local Rademacher complexity of $\mathcal{L}$
$$\mathcal{R}(\mathcal{L}^{\mathbf{r}}) := \mathcal{R}\Big\{\mathbf{a}\ell_{\mathbf{h}} \Big| \mathbf{a} \in [0,1], \ell_{\mathbf{h}} \in \mathcal{L}, \mathbf{L}[(\mathbf{a}\ell_{\mathbf{h}})^2] \leq \mathbf{r}\Big\}.$$

Sharper Generalization Bounds

- **Local Rademacher complexity of multi-class classification**
  With probability at least $1 - \delta$,
$$\mathcal{R}(\mathcal{L}^{\mathbf{r}}) \leq \frac{\mathbf{c}_{\mathbf{d},\vartheta}\xi(\mathbf{K})\sqrt{\zeta\mathbf{r}}\log^{\frac{3}{2}}(\mathbf{n})}{\sqrt{\mathbf{n}}} + \frac{4\log(1/\delta)}{\mathbf{n}},$$
  where
$$\xi(\mathbf{K}) = \begin{cases} \sqrt{\mathbf{e}}(4\log\mathbf{K})^{1+\frac{1}{2\log\mathbf{K}}}, & \text{if } \mathbf{q} \geq 2\log\mathbf{K}, \\ (2\mathbf{q})^{1+\frac{1}{\mathbf{q}}}\mathbf{K}^{\frac{1}{\mathbf{q}}}, & \text{otherwise}, \end{cases}$$
  $\mathbf{c}_{\mathbf{d},\vartheta}$ is a constant depending on $\mathbf{d}$ and ϑ.
- **A Sharper Generalization Bound**
  $\forall \mathbf{h} \in \mathcal{H}_{\mathbf{p},\kappa}$ and $\forall \mathbf{k} > \max(1, \frac{\sqrt{2}}{2\mathbf{d}})$, with probability at least $1 - \delta$,
$$\mathbf{L}(\mathbf{h}) \leq \max\left\{\frac{\mathbf{k}}{\mathbf{k}-1}\hat{\mathbf{L}}(\ell_{\mathbf{h}}), \hat{\mathbf{L}}(\ell_{\mathbf{h}}) + \frac{\mathbf{c}_{\mathbf{d},\vartheta,\zeta,\mathbf{k}}\xi^2(\mathbf{K})\log^3\mathbf{n}}{\mathbf{n}} + \frac{\mathbf{c}_\delta}{\mathbf{n}}\right\},$$
  where
$$\xi(\mathbf{K}) = \begin{cases} \sqrt{\mathbf{e}}(4\log\mathbf{K})^{1+\frac{1}{2\log\mathbf{K}}}, & \text{if } \mathbf{q} \geq 2\log\mathbf{K}, \\ (2\mathbf{q})^{1+\frac{1}{\mathbf{q}}}\mathbf{K}^{\frac{1}{\mathbf{q}}}, & \text{otherwise}, \end{cases}$$
  const $\mathbf{c}_{\mathbf{d},\vartheta}$ depends on $\mathbf{d}, \vartheta, \zeta, \mathbf{k}$, and const $\mathbf{c}_\delta$ depends δ.

Multi-Class Multiple Kernel Learning

- **Conv-MKL**
  Consider use multiple kernels $\kappa_\mu = \sum_{\mathbf{m}=1}^{\mathbf{M}} \mu_{\mathbf{m}}\kappa_{\mathbf{m}}$, the $\ell_{\mathbf{p}}$ hypothesis space of multiple kernels can be written as:
$$\mathcal{H}_{\mathrm{mkl}} = \Big\{\mathbf{h}_{\mathbf{w},\kappa_\mu} = (\langle \mathbf{w}_1, \phi_\mu(\mathbf{x})\rangle, \dots, \langle \mathbf{w}_{\mathbf{K}}, \phi_\mu(\mathbf{x})\rangle), \|\mathbf{w}\|_{2,\mathbf{p}} \leq 1, 1 < \mathbf{p} \leq \frac{2\log\mathbf{K}}{2\log\mathbf{K}-1}\Big\}.$$
  According to theoretical analysis, we add local Rademacher complexity (the tail sum of the eigenvalues of the kernel) to restrict $\mathcal{H}_{\mathrm{mkl}}$:
$$\mathcal{H}_2 = \Big\{\mathbf{h}_{\mathbf{w},\kappa_\mu} \in \mathcal{H}_{\mathrm{mkl}} : \sum_{\mathbf{m}=1}^{\mathbf{M}} \mu_{\mathbf{m}} \sum_{\mathbf{j}>\zeta} \lambda_{\mathbf{j}}(\mathbf{K}_{\mathbf{m}}) \leq 1\Big\}.$$
  Using normalized kernels $\tilde{\kappa}_{\mathbf{m}} = \Big(\sum_{\mathbf{j}>\zeta}\lambda_{\mathbf{j}}(\mathbf{K}_{\mathbf{m}})\Big)^{-1}\kappa_{\mathbf{m}}$ and $\tilde{\kappa}_\mu = \sum_{\mathbf{m}=1}^{\mathbf{M}} \mu_{\mathbf{m}}\tilde{\kappa}_{\mathbf{m}}$, we can simply rewrite $\mathcal{H}_2$ as
$$\Big\{\mathbf{h}_{\mathbf{w},\tilde{\kappa}_\mu} = (\langle \mathbf{w}_1, \tilde{\phi}_\mu(\mathbf{x})\rangle, \dots, \langle \mathbf{w}_{\mathbf{K}}, \tilde{\phi}_\mu(\mathbf{x})\rangle), \|\mathbf{w}\|_{2,\mathbf{p}} \leq 1, 1 < \mathbf{p} \leq \frac{2\log\mathbf{K}}{2\log\mathbf{K}-1}, \mu \succeq 0, \|\mu\|_1 \leq 1\Big\},$$
  With precomputed kernel matrices regularized by local Rademacher complexity, the method gets solution by any $\ell_{\mathbf{p}}$-norm MC-MKL solvers.
- **SMSD-MKL**
  Considering a more challenging case, we perform penalized ERM over the class $\mathcal{H}_1$, aiming to solve a convex optimization problem with an additional term representing local Rademacher complexity :
$$\min_{\mathbf{w},\mu} \underbrace{\frac{1}{\mathbf{n}}\sum_{\mathbf{i}=1}^{\mathbf{n}} \ell(\mathbf{w}, \phi_\mu(\mathbf{x}_{\mathbf{i}}), \mathbf{y}_{\mathbf{i}})}_{\mathbf{C}(\mathbf{w})} + \underbrace{\frac{\alpha}{2}\|\mathbf{w}\|_{2,\mathbf{p}}^2 + \beta\sum_{\mathbf{m}=1}^{\mathbf{M}} \mu_{\mathbf{m}}\mathbf{r}_{\mathbf{m}}}_{\Omega(\mathbf{w})},$$
  where
$$\ell(\mathbf{w}, \phi_\mu(\mathbf{x}_{\mathbf{i}}), \mathbf{y}_{\mathbf{i}}) = \Big|1 - \Big(\langle \mathbf{w}_{\mathbf{y}_{\mathbf{i}}}, \phi_\mu(\mathbf{x}_{\mathbf{i}})\rangle - \max_{\mathbf{y}\neq\mathbf{y}_{\mathbf{i}}}\langle \mathbf{w}_{\mathbf{y}}, \phi_\mu(\mathbf{x}_{\mathbf{i}})\rangle\Big)\Big|_+ \text{ and}$$
  $\mathbf{r}_{\mathbf{m}} = \sum_{\mathbf{j}>\zeta}\lambda_{\mathbf{j}}(\mathbf{K}_{\mathbf{m}})$ is the tail sum of the eigenvalues of the $\mathbf{m}$-th kernel matrix, $\mathbf{m} = 1, \dots, \mathbf{M}$.
  Based on widely used stochastic mirror descent framework, we design `SMSD-MKL` algorithm, implemented by stochastic sub-gradient descent with updating dual weights, to solve above optimization objective.

Experiments

Table: Comparison of average test accuracies of our `Conv-MKL` and `SMSD-MKL` with the others. We bold the numbers of the best method and underline the numbers of the other methods which are not significantly worse than the best one.

	Conv-MKL	SMSD-MKL	LMC	One vs. One	One vs. Rest	GMNP	ℓ_1 MC-MKL	ℓ_2 MC-MKL	UFO-MKL
plant	77.14±2.25	**78.01±2.17**	70.12±2.96	75.83±2.69	75.17±2.68	75.42±3.64	77.60±2.63	75.49±2.48	76.77±2.42
psortPos	74.41±3.35	**76.23±3.39**	63.85±3.94	73.33±4.21	71.70±4.89	73.55±4.22	71.87±4.87	70.70±4.89	74.56±4.04
psortNeg	74.07±2.16	**74.66±1.90**	57.85±2.49	73.74±2.87	71.94±2.50	74.27±2.51	72.83±2.20	72.42±2.65	73.80±2.26
nonpl	**79.15±1.51**	78.69±1.58	75.16±1.48	77.78±1.52	77.49±1.53	78.35±1.46	77.89±1.79	77.95±1.64	78.07±1.56
sector	92.83±2.62	**93.39±0.70**	93.16±0.66	90.61±0.69	91.34±0.61	\	\	92.15±2.57	92.60±0.47
segment	96.79±0.91	**97.62±0.83**	95.07±1.11	97.08±0.61	97.02±0.80	96.87±0.80	96.98±0.64	97.58±0.68	97.20±0.82
vehicle	**79.35±2.27**	77.28±2.78	75.61±3.56	78.72±1.92	79.11±1.94	81.57±2.24	74.96±2.93	76.27±3.15	76.92±2.83
vowel	98.82±1.19	**98.83±5.57**	62.32±4.97	98.12±1.76	98.22±1.83	97.04±1.85	98.27±1.22	97.86±1.75	98.22±1.62
wine	**99.63±0.96**	**99.63±0.96**	97.87±2.80	97.24±3.05	98.14±3.04	97.69±2.43	98.61±1.75	98.52±1.89	99.44±1.13
dna	96.08±0.83	**96.30±0.79**	92.02±1.50	95.89±0.56	95.61±0.73	94.60±0.94	96.27±0.68	95.06±0.92	95.84±0.61
glass	**75.19±5.05**	73.72±5.80	63.95±6.04	71.98±5.75	70.00±5.75	71.24±8.14	69.07±8.08	74.03±6.41	72.46±6.12
iris	96.67±2.94	**97.00±2.63**	88.00±7.82	95.93±3.25	95.87±3.20	95.40±7.34	95.40±6.46	94.00±7.82	95.93±2.88
svmguide2	82.69±5.65	**85.17±3.83**	81.10±4.15	84.79±3.45	84.27±3.03	81.77±3.45	83.16±3.63	83.84±4.21	82.91±3.09
satimage	91.64±0.88	91.78±0.82	84.95±1.15	90.67±0.91	89.29±0.96	89.97±0.81	91.86±0.62	90.43±1.27	**91.92±0.83**

Conclusions

- A new local Rademacher complexity based bound with fast convergence rate for multi-class classification is establish. Convergence rate is improved from sub-linear to linear $\mathcal{O}\big(\sqrt{\frac{\mathbf{K}}{\mathbf{n}}}\big) \Rightarrow \mathcal{O}\big(\frac{(\log\mathbf{K})^{2+1/\log\mathbf{K}}}{\mathbf{n}}\big)$.
- Two novel multi-class classification algorithms are proposed with statistical guarantees: a) `Conv-MKL`. b) `SMSD-MKL`.

Main References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson.
Local Rademacher complexities.
The Annals of Statistics, 33(4):1497–1537, 2005.

[2] F. Orabona and J. Luo.
Ultra-fast optimization algorithm for sparse multi kernel learning.
Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pages 249–256, 2011.

https://lijian.ac.cn, https://iie-liuyong.github.io {lijian9026,liuyong}@iie.ac.cn