Improving the Generalization Performance of Multi-class SVM via Angular Regularization

Jianxin Li [1], Haoyi Zhou [1], Pengtao Xie [2,3], Yingchun Zhang [1]

[1] Beihang University, [2] Carnegie Mellon University, [3] Petuum Inc

IJCAI-17 MELBOURNE

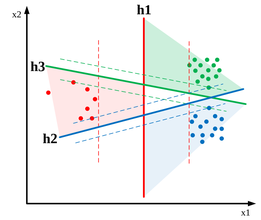
I. ABSTRACT

In multi-class support vector machine (MSVM) for classification, one core issue is to regularize the coefficient vectors to reduce overfitting. Various regularizers have been proposed such as L_2, L_1, and trace norm. In this paper, we introduce a new type of regularization approach -- angular regularization, that encourages the coefficient vectors to have larger angles such that class regions can be widen to flexibly accommodate unseen samples. We propose a novel angular regularizer based on the singular values of the coefficient matrix, where the uniformity of singular values reduces the correlation among different classes and drives the angles between coefficient vectors to increase. In generalization error analysis, we show that decreasing this regularizer effectively reduces generalization error bound. On various datasets, we demonstrate the efficacy of the regularizer in reducing overfitting.

II. OVERVIEW

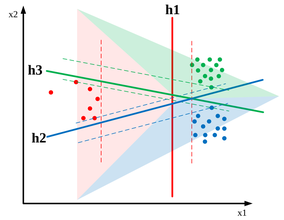
Multi-class SVM classification ($K > 2$ classes):

O-V-A Ensemble



- One-Against-ALL
- One-Against-One
- Directed acyclic graph SVM

Crammer Formulation



- Crammer and Singer (2001)
- Analyzable formulation
- Multiple kernelization

Multi-class SVM formulation:

$$\min_{\mathbf{W}} P(\mathbf{W}) \triangleq \min_{\mathbf{W}} \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{W};(\mathbf{x}_i,y_i))$$

$$L(\mathbf{W};(\mathbf{x}_i,y_i)) = \max(0,\ 1+\mathbf{w}_{R_i}^T\cdot\mathbf{x}_i-\mathbf{w}_{y_i}^T\cdot\mathbf{x}_i),$$
$$\text{where } \mathrm{R_i} = \underset{k\in\mathcal{Y},k\neq y_i}{\operatorname{argmax}}\ \mathbf{w}_k^T\cdot\mathbf{x}_i.$$

Multi-class SVM with norm-based Regularization:

- L1-norm (Bradley et al, 1998; Wang et al, 2012),

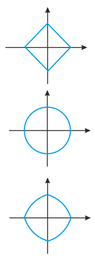
$$\min_{\mathbf{W}} \lambda\|\mathbf{W}\|_1 + \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{W};(\mathbf{x}_i,y_i))$$

- L2-norm (Weston et al, 1999; Guermuer, 2002),

$$\min_{\mathbf{W}} \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{W};(\mathbf{x}_i,y_i))$$

- Doubly-norm (Wang et al, 2006)

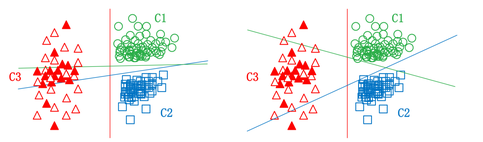
$$\min_{\mathbf{W}} \lambda_1\|\mathbf{W}\|_1 + \frac{\lambda_2}{2}\|\mathbf{W}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{W};(\mathbf{x}_i,y_i))$$



Angular perspective to reduce Overfitting:

Table 1: Classification Accuracy of MSVM-ℓ_2

Dataset	Yeast	Usps	YaleB
Dim. of Features	8	256	1024
Train Accuracy (%)	57.23	97.31	97.70
Test Accuracy (%)	52.00	90.48	93.25



(a) Unregularized MSVM (b) Angular regularized MSVM

III. METHODS

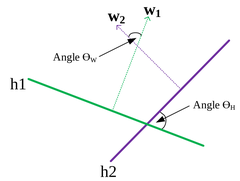
Desirable diversity measure:

INVARIANT to scale, translation, rotation and orientation of the two hyperplanes.

$$\Theta_{\mathrm{H}} = \Theta_{\mathrm{W}}$$

Angle of hyperplanes — Angle of weight vectors

$$\theta_{ij} = \arccos\left(\frac{|\mathbf{w}_i^\top\mathbf{w}_j|}{\|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2}\right)$$



Angular Regularizer (AR):

The negative of the minimum angle:

$$\mathcal{R}(\mathbf{W}) = -\min_{i\neq j}\theta_{ij}$$

- Pros: directly encourages theses coefficient vectors to have larger angles, and this definition facilitates theoretical analysis.
- Cons: non-smooth (similar to the one in Xie et al, 2015).

Decorrelation Regularizer (DR):

Given two hyperplanes ($\mathrm{h_i}$ and $\mathrm{h_j}$), we can treat i and j as two random variables:

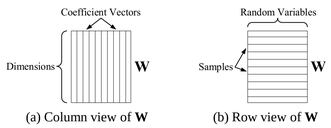
$\mathrm{h_i}$ — Samples of i: $\mathbf{W}_\mathrm{i} \leftarrow \{w_{in}\}_{n=1}^D$

$\mathrm{h_j}$ — Samples of j: $\mathbf{W}_\mathrm{j} \leftarrow \{w_{jn}\}_{n=1}^D$

$$\rho_{ij} = \frac{\sum_{n=1}^D (w_{in}-\bar{w}_i)(w_{jn}-\bar{w}_j)}{\sqrt{\sum_{n=1}^D (w_{in}-\bar{w}_i)^2\sum_{n=1}^D (w_{jn}-\bar{w}_j)^2}} \quad\Rightarrow\quad \rho_{ij} = \cos(\theta_{ij})$$

$$\bar{w}_i = \tfrac{1}{n}\textstyle\sum_{n=1}^D w_{in} = 0 \qquad \bar{w}_j = \tfrac{1}{n}\textstyle\sum_{n=1}^D w_{jn} = 0$$

A set of hyperplanes' weights vectors $\mathbf{W}$ can be reformulated as full-rank matrix with rank $K < D$: $\mathbf{W}=[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K]$.



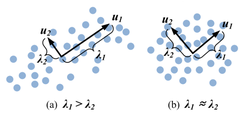
(a) Column view of $\mathbf{W}$ (b) Row view of $\mathbf{W}$

We compute the full-rank $K \times K$ Gram matrix $\mathbf{G} = \mathbf{W}^\top\mathbf{W}$, and

$$\mathbf{G} = \textstyle\sum_{i=1}^K \lambda_i \mathbf{u}_i\mathbf{u}_i^\top$$

Eigenvector $\mathbf{u}_i$ of $\mathbf{G}$ represents a principal direction of the point cloud.

The associated eigenvalue λ_i measures the variability of points along that direction.



(a) $\lambda_1 > \lambda_2$ (b) $\lambda_1 \approx \lambda_2$

We define a probability distribution of the eigenvalues

$$p(X=i) = \frac{\lambda_i}{\sum_{j=1}^K \lambda_j} \qquad \text{(normalization)}$$

If $p(X)$ is close to uniform distribution $q(X=i) = \frac{1}{K}$, hyperplanes diversified. And we measure it by Kullback-Leibler divergence:

$$\begin{aligned} KL(q||p) &= \textstyle\sum_{i=1}^K \frac{1}{K}\log\frac{1/K}{\lambda_i/\sum_{j=1}^K\lambda_j} \\ &= \log\textstyle\sum_{j=1}^K\lambda_j - \frac{1}{K}\sum_{i=1}^K\log\lambda_i - \log K \end{aligned}$$

Using two facts $\mathrm{tr}(\mathbf{G}) = \sum_{j=1}^K\lambda_j$ and $\det(\mathbf{G}) = \prod_{i=1}^K\lambda_i$, we have

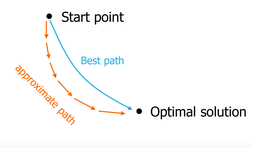
$$KL(q||p) = \log\mathrm{tr}(\mathbf{W}^\top\mathbf{W}) - \tfrac{1}{K}\log\det(\mathbf{W}^\top\mathbf{W}) - \log K$$

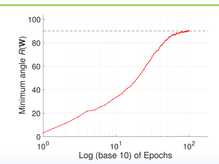
Dropped

The Decorrelation Regularizer is

$$\widehat{\mathcal{R}}(\mathbf{W}) = \log\mathrm{tr}(\mathbf{W}^\top\mathbf{W}) - \tfrac{1}{K}\log\det(\mathbf{W}^\top\mathbf{W})$$

Theorem 1. *Let $\nabla\widehat{\mathcal{R}}$ be the gradient of $\widehat{\mathcal{R}}(\mathbf{W})$ w.r.t $\mathbf{W}$. $\exists\kappa > 0$, such that $\forall\,\eta\in(0,\kappa)$, $\mathcal{R}(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}}) \le \mathcal{R}(\mathbf{W})$.*





IV. OPTIMIZATION

An Angle-regularized MSVM (AR-MSVM)

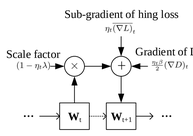
MSVM-L2

$$\min_{\mathbf{W}} \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{W};(\mathbf{x}_i,y_i)) + \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \frac{\beta}{2}\mathcal{R}(\mathbf{W})$$

Objective function for Optimization

$$\min_{\mathbf{W}} \frac{1}{m}\sum_{i=1}^{m}\max(0,\ 1+\mathbf{w}_{r_i}^\top\mathbf{x}_i-\mathbf{w}_{y_i}^\top\mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \frac{\beta}{2}\widehat{\mathcal{R}}(\mathbf{W})$$

Optimizing method: Pegasos-style Stochastic sub-gradient method (Shalev-Shwartz et al., 2011; Wang et al., 2010).



- Learning rate η_t is updated with ADADELTA (Zeiler, 2012)
- For each hyperplane k, the sub-gradient of its hing loss is

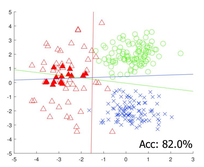
$$(\nabla L)_t^k = \begin{cases} -\mathbf{x}, & if\ k = y \\ +\mathbf{x}, & if\ k = R \\ 0. & \text{otherwise} \end{cases}$$

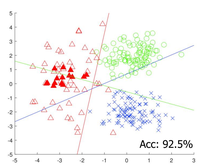
- The gradient of DR is

$$\nabla\widehat{\mathcal{R}} = \frac{2\mathbf{W}}{\mathrm{tr}(\mathbf{W}^\top\mathbf{W})} - \frac{2}{K}\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}$$

V. EXPERIMENTAL RESULTS

Case study: Intuitive experiments





Multi-class Classification

Datasets:

Table 2: Statistics of Datasets

Dataset	#Classes	#Train	#Test	#Features
YaleB	38	1500	914	1024
ImageNet-50	50	30K	10K	128
Covtype	7	100K	40K	54
Shuttle	7	30450	14500	9
New-thyroid	3	108	107	5
Yeast	10	1134	350	8
Dermatology	6	323	35	33
Page-Blocks	5	4924	548	10
Wine-Quality-Red	6	1439	160	11
Zoo	7	89	12	16

- 10 random train/test splits
- Parameters is tuned via 5-fold CV
- Results is averaged over 10 runs

Baseline:

- C-SVC (Chang and Lin, 2011): performing multi-class classification using an O-A-O strategy.
- AMM (Wang et al, 2011): a multi-class classification method based on adaptive multi-hyperplanes.
- MSVM-L1 (Shalev-Shwartz et al, 2011): L1-regularized MSVM.
- CS-MSVM (Yu et al, 2011): MSVM regularized by a cosine similarity regularizer.
- IC-MSVM (Bao et al, 2013): MSVM regularized by an incoherence regularizer.
- Div-MSVM (Xie et al, 2015): MSVM regularized by a diversity-promoting regularizer.
- MSVM-Struct (Tsochantaridis et al, 2004): L2-regularized MSVM with a cutting-plane algorithm.
- MSVM-L2 (Wang et al, 2010): L2-regularized MSVM solved with a SGD algorithm.

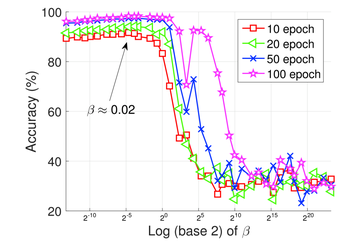
Table 3: Classification results (%) on six multi-class datasets

Dataset	Metric	C-SVC	AMM	MSVM-ℓ_1	CS-MSVM	IC-MSVM	Div-MSVM	MSVM-Struct	MSVM-ℓ_2	AR-MSVM
YaleB	*Acc*	55.31 ±0.1	83.32 ±2.5	85.12 ±4.4	94.13 ±1.6	89.38 ±0.9	94.25 ±0.4	91.50 ±0.0	93.90 ±0.5	**94.55** ±0.8
	F	54.74 ±1.0	86.88 ±1.8	87.11 ±2.6	93.71 ±1.3	90.54 ±1.4	93.87 ±0.3	91.39 ±0.0	93.51 ±0.5	**94.38** ±0.7
ImageNet-50	*Acc*	91.59 ±0.4	92.27 ±0.1	91.23 ±0.0	92.86 ±0.1	92.96 ±0.1	92.93 ±0.1	89.36 ±0.0	92.32 ±0.1	**93.12** ±0.1
	F	91.70 ±0.4	92.31 ±0.1	91.38 ±0.0	92.90 ±0.1	92.98 ±0.1	92.97 ±0.1	89.65 ±0.0	92.35 ±0.0	**93.17** ±0.1
Covtype	*Acc*	71.35 ±0.5	70.21 ±0.3	56.12 ±0.4	70.29 ±1.1	70.82 ±0.2	70.70 ±0.1	65.84 ±0.1	69.45 ±0.2	**71.75** ±0.1
	F	50.97 ±3.1	49.84 ±2.4	37.03 ±3.5	50.67 ±2.1	50.82 ±0.1	50.53 ±0.5	36.80 ±0.1	45.64 ±1.8	**51.31** ±0.9
Shuttle	*Acc*	**98.86** ±0.3	93.81 ±0.3	79.41 ±0.4	96.02 ±0.0	96.22 ±0.1	94.74 ±0.2	66.57 ±0.0	92.83 ±0.1	97.08 ±0.3
	F	52.94 ±2.1	54.12 ±0.3	24.90 ±0.0	57.20 ±0.1	57.57 ±0.1	55.61 ±0.2	33.66 ±0.0	53.55 ±0.2	**58.53** ±0.3
New-thyroid	*Acc*	76.92 ±0.9	86.77 ±1.8	86.15 ±0.0	78.46 ±10.9	78.69 ±3.3	91.54 ±0.8	89.23 ±0.0	90.87 ±0.0	**92.15** ±0.5
	F	58.75 ±1.0	83.66 ±2.4	87.50 ±0.0	75.32 ±2.1	78.46 ±2.1	90.08 ±1.1	87.50 ±0.0	89.24 ±0.0	**90.91** ±0.7
Yeast	*Acc*	**59.71** ±1.5	48.63 ±5.3	53.71 ±2.4	53.61 ±6.2	54.14 ±0.2	53.14 ±2.5	49.93 ±0.1	51.86 ±3.8	54.80 ±1.1
	F	**61.81** ±2.0	44.87 ±5.1	48.63 ±4.8	52.80 ±4.1	52.07 ±2.0	52.64 ±3.7	37.09 ±0.1	49.45 ±3.1	54.65 ±4.3

Angular regularizer can improve the performance of MSVM.

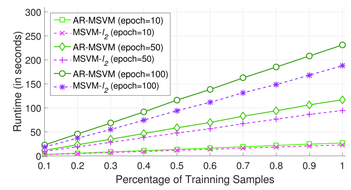
- Ensemble methods: efficiently improve the generalization performance.
- L1-regularized methods: sparse constraints are not effective in reduce overfitting of multi-class classification.
- Diversity-promoting methods: CS-MSVM is smooth, while the others are non-smooth. AR-MSVM is not only smooth but also ensures all angles are increased directly.
- L2-regularized methods: AR-MSVM is directly based on MSVM-L2, but achieves better results in all cases.

Parameter Sensitivity:



- 2^{-10} to 2^{-2} : a larger β results in larger angles
- 2^{-2} to 2^{20} : too large β harms hing loss

Runtime:



Linux Machine
CPU: 2.00GHz Xeon CPU
Memory: 256 GB
Hard Drive: 512 GB
magnetic disk

Imbalanced Classification

Table 4: Classification accuracy (%) on four imbalance datasets

Dataset	Metric	Global-CS	CS	Static-SMT	SL-SMT	AdaB.NC	MSVM-ℓ_2	AR-MSVM	Δ
Dermatology	*Acc*	95.78	95.44	95.60	94.30	**97.08**	84.29 ± 8.959	94.57 ± 1.622	+10.28
Page-Blocks	*Acc*	91.67	89.32	69.04	89.34	88.29	88.93 ± 2.237	**92.27** ± 1.019	+3.34
Wine-Quality-Red	*Acc*	39.33	37.82	30.74	37.93	37.22	49.94 ± 5.064	**52.94** ± 1.793	+3.00
Zoo	*Acc*	95.02	93.02	95.35	93.02	95.02	93.28 ± 2.500	**95.38** ± 1.510	+2.10

A benchmark established by FernaNdez et al, 2013

Table 5: *Acc* (%) on each class in the Page-Blocks dataset

Class	1	2	3	4	5	sum
#Train	4421	296	25	79	103	4924
#Test	492	33	3	8	12	548
MSVM-ℓ_2	96.75	21.21	66.67	12.50	16.67	89.05
AR-MSVM	95.93	54.55	66.67	50.00	75.00	92.15

- On small classes (2-5), AR-MSVM performs much better than MSVM.
- On large class (1), they are comparable.

VI. THEORETICAL RESULTS

We analyze the generalization error under a supervised learning framework via the PAC learning theory.

Theorem 1. *Fix $p > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following multi-class classification genralization bound holds for all $h \in H \rightsquigarrow H^1$:*

$$L(\hat{h}) - L(h^*) \le \tfrac{8K^2C_1C_2}{p\sqrt{m}} + \tfrac{2}{\sqrt{m}} + (\tfrac{1}{p}\sqrt{(1-\tfrac{1}{K})(\cos(-\mathcal{R}(\mathbf{W})) + \tfrac{K+1}{K-1})}C_1C_2 + 1)\sqrt{\tfrac{2\log(2/\delta)}{m}}.$$

Main conclusions:

- The MSVM's objective function is seeking the maximum margin principle which encourages H approaching H^1.
- The generalization error bound $L(\hat{h}) - L(h^*)$ is an increasing function of the angular regularizer $\mathcal{R}(\mathbf{W})$. It is also why β needs to be chosen to achieve the best accuracy.
- Thus, decreasing angular regularizer $\mathcal{R}(\mathbf{W})$ can reduce the error bound and improve the generalization performance.

ACKNOWLEDGEMENTS

This work is supported by China 973 Fundamental R&D Program (No.2014CB340300), NSFC program (No.61472022, 61421003), SKLSDE-2016ZX-11, and partly by the Beijing Advanced Innovation Center for Big Data and Brain Computing.

Appendix Files — Wechat Contact — Homepage