An Unbiased Risk Estimator for Learning with Augmented Classes

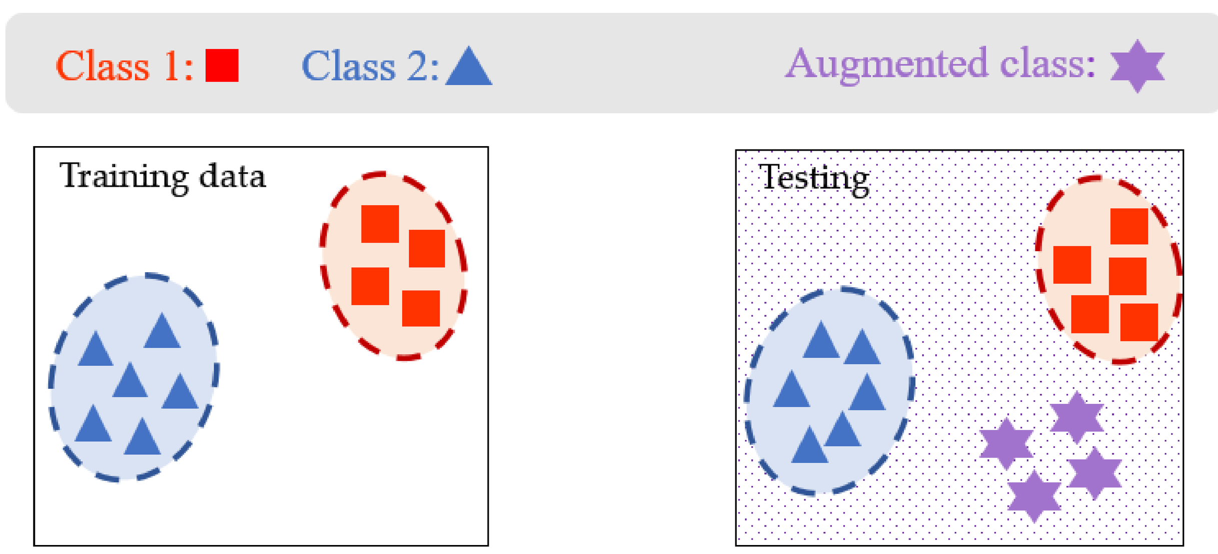
Yu-Jie Zhang, Peng Zhao, Lanjihong Ma, Zhi-Hua Zhou

Contact

{zhangyj, zhaop, maljh, zhouzh}@lamda.nju.edu.cn

Learning with Augmented Classes

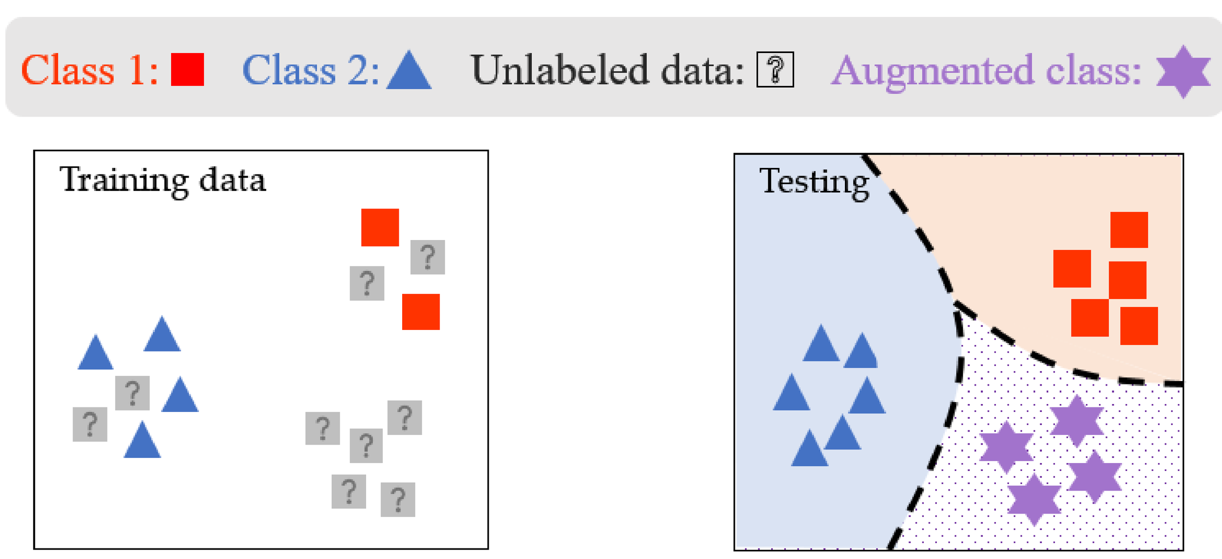
We study the **learning with augmented classes** problem (LAC), where augmented classes unobserved in training data might emerge in the testing phase.



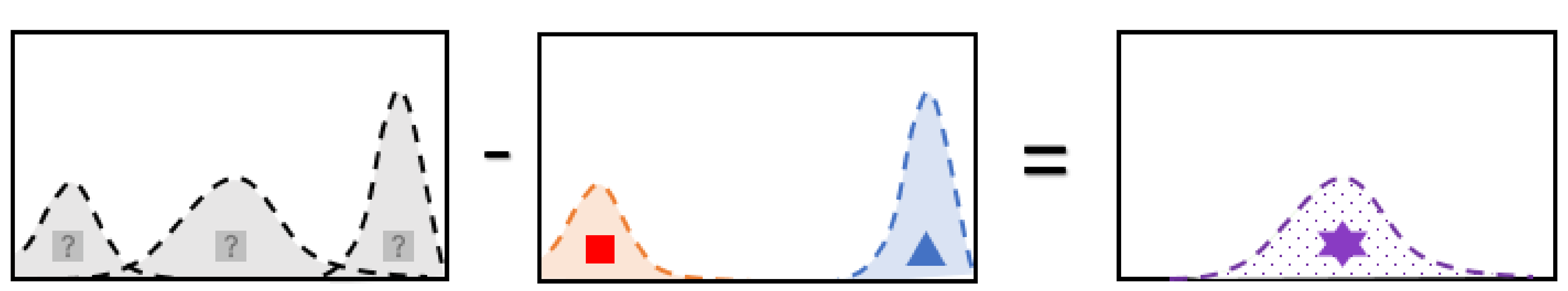
- Previous studies generally attempt to discover augmented classes by exploiting their *geometric properties*
- ***Generalization ability*** of learned models is less explored

Exploiting Unlabeled Data

By exploiting ***unlabeled data***, we develop the EULAC approach, which enjoys ***sound theoretical guarantees***.



Why unlabeled data is helpful?



Intuition: approximate the distribution of augmented classes by separating the distribution of known classes from unlabeled data

EULAC Approach

Class shift condition: testing distribution P_{te} is ***a mixture*** of those of known P_{kc} and augmented classes P_{ac}.

$$P_{te} = \theta \cdot P_{\text{kc}} + (1 - \theta) \cdot P_{\text{ac}}$$

Where $\boldsymbol{\theta} \in [0,1]$ is the mixture proportion.

Equivalence of the risk: under the class shift condition,

$$R_\psi = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}} \left[\Psi(\boldsymbol{f}(\mathbf{x}), y)\right]$$

Classifiers' risk over testing distribution

equal

$$R_{LAC} = \theta \cdot \mathbb{E}_{(\mathbf{x},y)\sim P_{\text{kc}}} \left[\Psi(\boldsymbol{f}(\mathbf{x}), y)\right] + \mathbb{E}_{\mathbf{x}\sim p_X^{te}(\mathbf{x})} \left[\Psi(\boldsymbol{f}(\mathbf{x}), \text{ac})\right]$$

LAC risk R_{LAC} can be assessed in training with labeled data and unlabeled data. Different algorithms can be derived by minimizing R_{LAC} on various hypothesis space.

Generalized class shift condition: consider the distribution change on known classes together with augmented classes

$$P_{te} = \underbrace{\theta_{te}^1 \cdot P_1 + \theta_{te}^2 P_2 + \cdots}_{\theta \cdot \widetilde{P}_{\text{kc}}} + (1 - \textstyle\sum_{k=1}^K \theta_{te}^k) \cdot P_{\text{ac}}$$

prior change happens on known classes

Equivalence $R_{LAC} = R_\psi$ also holds under the generalized condition

Algorithms

- ☐ Empirical risk minimization on ***kernel-based hypothesis set***

$$\min_{f_1,\ldots,f_K,f_{\text{ac}}\in\mathbb{F}} \widehat{R}_{LAC} + \lambda\Big(\sum_{k=1}^K \|f_k\|_{\mathbb{F}}^2 + \|f_{\text{ac}}\|_{\mathbb{F}}^2\Big)$$

The optimization problem is ***convex*** if we choose:

- multiclass loss function Ψ as ***one-vs-rest loss (OVR)***
- binary loss function ψ in OVR loss is convex and satisfies

$$\psi(z) - \psi(z) = -z \text{ for all } z \in \mathbb{R}$$

- ☐ We also minimize R_{LAC} with ***deep models***

Theoretical Analysis

Theorem 2 (Infinite-sample Consistency). *Under the same condition with Theorem 1, when using $\psi(z) = (1-z)^2/4$ as the surrogate loss function, we have*

$$R(f) - R^* \le \sqrt{2\big(R_{LAC}(f_1,\ldots,f_K,f_{nc}) - R_{LAC}^*\big)},$$

which holds for all measurable functions $f_1,\ldots,f_K,f_{nc}$ and $f(\mathbf{x}) = \arg\max_{k\in\{1,\ldots,K,nc\}} f_k(\mathbf{x})$. Here, $R_{LAC}^* = \min_{f_1,\ldots,f_K,f_{nc}} R_{LAC}(f_1,\ldots,f_K,f_{nc})$ and $R^* = \min_f R(f) = \mathbb{E}_{(\mathbf{x},y)\sim P_{te}}[\mathbb{1}(f(\mathbf{x}) \neq y)]$ is the Bayes error over the testing distribution.

Consistency: minimizing R_{LAC} *is identical to* minimize the 0-1 risk R

Theorem 4 (Finite-sample Convergence). *Under assumptions of Theorem 3 and let $\widehat{f}_1,\ldots,\widehat{f}_K,\widehat{f}_{nc}$ be the optimal solution of the optimization problem (8) with certain $\lambda > 0$, we have*

$$R_{LAC}(\widehat{f}_1,\ldots,\widehat{f}_K,\widehat{f}_{nc}) - \inf_{\boldsymbol{f}\in\mathscr{F}} R_{LAC}(f_1,\ldots,f_K,f_{nc}) \le \mathcal{O}\left(\frac{K+1}{\sqrt{n_l}} + \frac{K+1}{\sqrt{n_u}}\right),$$

where $\boldsymbol{f}$ denotes $(f_1,\ldots,f_K,f_{nc})$ and $\mathscr{F} = \{\boldsymbol{f} \mid f_1,\ldots,f_K,f_{nc} \in \mathbb{F}, \sum_{k=1}^K \|f_k\|_{\mathbb{F}}^2 + \|f_{nc}\|_{\mathbb{F}}^2 \le c_\lambda^2\}$. The parameter $c_\lambda > 0$ is a constant related to λ in (8). We use the $\mathcal{O}$-notation to keep the dependence on n_u, n_l and K only.

Convergence: our kernel-based approach can minimize R_{LAC}

Experiments

Table 1:MacroF1 comparison on 10 benchmark datasets

Dataset	OVR-SVM	W-SVM	OSNN	EVM	LACU-SVM	PAC-iForest	EULAC
usps	75.42 ± 4.87 •	79.77 ± 4.97 •	63.14 ± 8.91 •	61.14 ± 6.27 •	69.20 ± 8.34 •	55.69 ± 13.3 •	**86.52 ± 2.72**
segment	71.78 ± 5.12 •	80.82 ± 9.38 •	85.10 ± 5.98	82.13 ± 5.88 •	40.69 ± 12.5 •	63.64 ± 13.1 •	**86.17 ± 5.80**
satimage	54.67 ± 9.80 •	76.29 ± 13.2 •	62.48 ± 11.2 •	72.10 ± 8.16 •	51.56 ± 17.3 •	60.76 ± 7.79 •	**81.25 ± 6.18**
optdigits	80.11 ± 3.80 •	87.82 ± 4.64 •	86.97 ± 3.79 •	72.00 ± 8.33 •	80.92 ± 3.68 •	71.65 ± 5.46 •	**91.54 ± 2.95**
pendigits	72.78 ± 5.19 •	87.79 ± 3.95	86.69 ± 3.39 •	**89.94 ± 1.30**	70.66 ± 6.18 •	73.21 ± 4.52 •	88.41 ± 4.81
SenseVeh	48.07 ± 3.80 •	45.96 ± 2.32 •	49.91 ± 6.88 •	51.24 ± 3.91 •	51.61 ± 3.31 •	54.12 ± 7.19 •	**77.33 ± 2.17**
landset	60.43 ± 7.65 •	68.91 ± 17.0 •	73.25 ± 9.23 •	76.00 ± 7.79 •	53.59 ± 9.88 •	70.50 ± 7.16 •	**85.70 ± 4.46**
mnist	66.74 ± 2.76 •	75.38 ± 4.62 •	57.75 ± 10.9 •	58.39 ± 5.94 •	63.53 ± 7.58 •	48.31 ± 9.62 •	**80.66 ± 5.38**
shuttle	37.39 ± 14.1 •	58.48 ± 34.5 •	48.21 ± 16.4 •	–	34.18 ± 13.4 •	29.36 ± 8.70 •	**66.49 ± 17.9**
EULAC w/ t/ l	9/ 0/ 0	8/ 1/ 0	8/ 1/ 0	8/ 1/ 0	9/ 0/ 0	9/ 0/ 0	rank first 8/ 9

Figure 2: growing performance with more unlabeled data

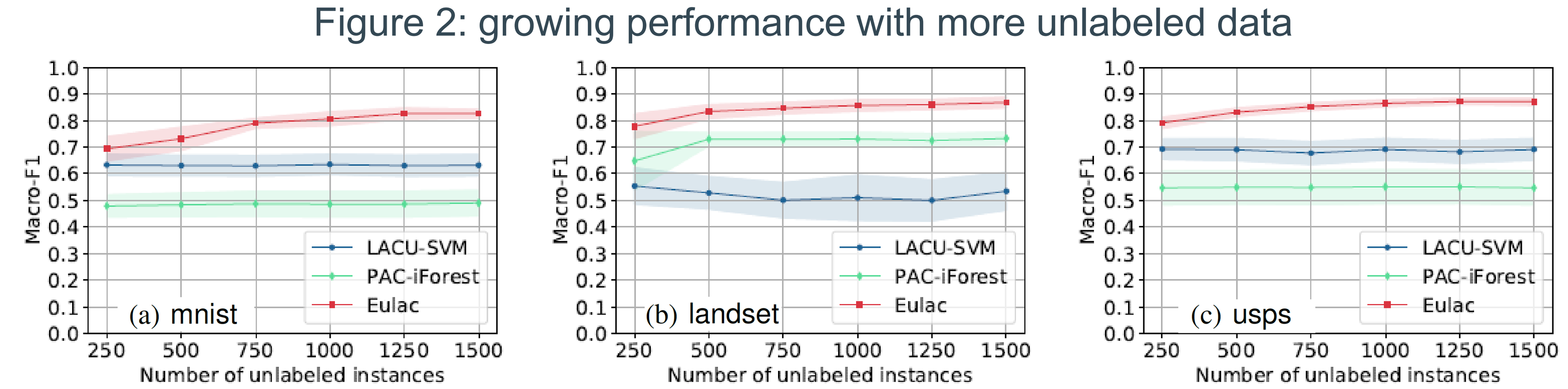
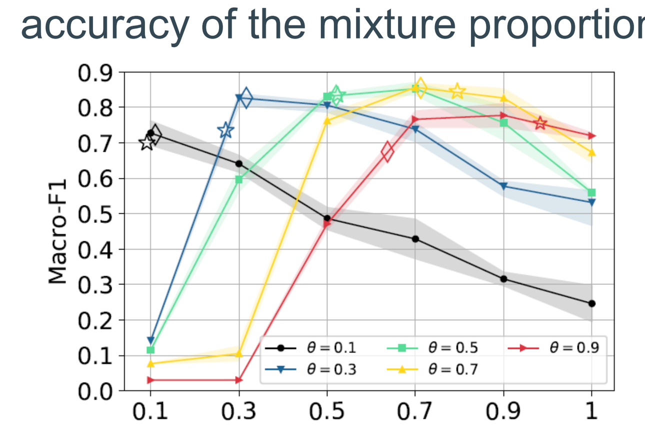


Figure 3: Influence and estimation accuracy of the mixture proportion



Results

- EULAC ranks first on ***8/9 datasets***
- EULAC uses ***unlabeled*** data well
- Performance will be influenced by θ, but ***its estimation is accurate***