Noname manuscript No.
(will be inserted by the editor)

Supplemental Material

ARConvL: Adaptive Region-Based Convolutional Learning for Multi-class Imbalance Classification

Submission ID: 897

the date of receipt and acceptance should be inserted later

Abstract This document presents additional information of the submitted paper "ARConvL: Adaptive Region-Based Convolutional Learning for Multi-class Imbalance Classification". The information of datasets, parameter setting of investigated methods, and experimental results in terms of class-wise accuracy are presented.

Address(es) of author(s) should be given

1 Datasets

Table 1 Information of datasets

Data	Classes	IR	Training Set	Testing Set
Mnist-1	10	1.2	5923: 6742: 5958: 6131: 5842: 5421: 5918: 6265: 5851: 5949	980: 1135: 1032: 1010: 982: 892: 958: 1028: 974: 1009
Mnist-10	10	11.5	592: 6742: 595: 6131: 584: 5421: 591: 6265: 585: 5949	
Mnist-20	10	23.1	296: 6742: 297: 6131: 292: 5421: 295: 6265: 292: 5949	
Mnist-50	10	58.1	118: 6742: 119: 6131: 116: 5421: 118: 6265: 117: 5949	
Mnist-100	10	116.2	59: 6742: 59: 6131: 58: 5421: 59: 6265: 58: 5949	
Fashion-1	10	1.0	6000: 6000: 6000: 6000: 6000: 6000: 6000: 6000: 6000: 6000	1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000
Fashion-10	10	10.0	600: 6000: 600: 6000: 600: 6000: 600: 6000: 600: 6000	
Fashion-20	10	20.0	300: 6000: 300: 6000: 300: 6000: 300: 6000: 300: 6000	
Fashion-50	10	50.0	120: 6000: 120: 6000: 120: 6000: 120: 6000: 120: 6000	
Fashion-100	10	100.0	60: 6000: 60: 6000: 60: 6000: 60: 6000: 60: 6000	
SVHN-1	10	3.0	13861: 10585: 8497: 7458: 6882: 5727: 5595: 5045: 4659: 4948	5099: 4149: 2882: 2523: 2384: 1977: 2019: 1660: 1595: 1744
SVHN-10	10	22.8	1386: 10585: 849: 7458: 688: 5727: 559: 5045: 465: 4948	
SVHN-20	10	45.6	693: 10585: 424: 7458: 344: 5727: 279: 5045: 232: 4948	
SVHN-50	10	113.8	277: 10585: 169: 7458: 137: 5727: 111: 5045: 93: 4948	
SVHN-100	10	230.1	138: 10585: 84: 7458: 68: 5727: 55: 5045: 46: 4948	
Cifar10-1	10	1.0	5000: 5000: 5000: 5000: 5000: 5000: 5000: 5000: 5000: 5000	1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000: 1000
Cifar10-10	10	10.0	500: 5000: 500: 5000: 500: 5000: 500: 5000: 500: 5000	
Cifar10-20	10	20.0	250: 5000: 250: 5000: 250: 5000: 250: 5000: 250: 5000	
Cifar10-50	10	50.0	100: 5000: 100: 5000: 100: 5000: 100: 5000: 100: 5000	
Cifar10-100	10	100.0	50: 5000: 50: 5000: 50: 5000: 50: 5000: 50: 5000	
CelebA	5	10.5	24267: 33192: 6896: 3713: 38906	3056: 4793: 967: 411: 4144
iNaturalist 2018	14	8996.8	1959: 2789: 1492: 1009: 35987: 4: 155: 1716: 21798: 5026: 2002: 29700: 53: 5688	1107: 432: 534: 342: 3774: 3: 75: 963: 6093: 702: 786: 8751: 12: 852

2 Parameter Setting

Table 2 Parameter settings for deep methods investigated.

Methods	Parameters
CPL	/
GCPL	$\lambda = 0.1$
Focal	$\alpha = 2$
CB	$\beta = 0.9999$
CB Focal	$\beta = 0.999$, $\alpha = 0.5$ for iNaturalist 2018, $\beta = 0.9999$, $\alpha = 2$ for others
Affinity	$\lambda = 0.75$ for MNIST and Fashion MNIST, $\lambda = 0.43$ for others; $\sigma = 10$
LA	$\tau = 1$
ARConvL	$\gamma = 0.05$

3 Experimental Results

3.1 Performance Comparison

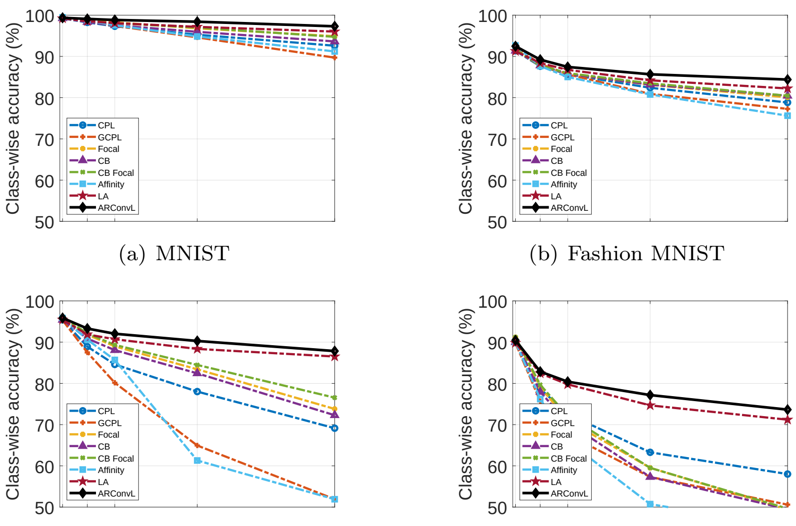
Table 3 shows that our ARConvL achieves the best class-wise accuracy in 16 out of 20 datasets, showing the effectiveness of our approach in dealing with varying levels of class imbalance. Friedman tests at the significance level 0.05 reject H0 with the p-value 0, meaning that there is significant difference between methods. The average rank of ARConvL is 1.4, being the best (lowest value) among all competing methods. This indicates that our method generally performs the best across datasets with different levels of class imbalance. ARConvL is then chosen as the control method to conduct post-hoc tests for performing the best among all classifiers. Post-hoc tests show that the proposed ARConvL significantly outperforms all competitors.

3.2 Performance Deterioration with Increasing Imbalance Levels

Figure 1 shows experimental results in terms of class-wise accuracy. We can see that all methods achieve similar class-wise accuracy in the original image repository for the case $q = 1$. With the increase of class imbalance levels with larger q, Performance of all methods declines. Yet, the proposed ARConvL can usually achieve better class-wise accuracy than its competitors when datasets become more imbalanced, demonstrating the robustness of ARConvL against different levels of class imbalance.

Table 3 Class-wise accuracy (%) of the investigated methods. Each entry is the mean±std of 10 times. The last column corresponds to our ARConvL. The best model on each dataset is highlighted in bold. The last row lists the average ranks (avgRank) of each model across datasets. Significant difference against ARConvL is highlighted in yellow.

Data	CPL	GCPL	Focal	CB	CB Focal	Affinity	LA	ARConvL
Mnist-1	99.2±0.1	99.4±0.0	99.4±0.1	99.2±0.1	99.4±0.1	**99.5±0.1**	99.2±0.1	99.4±0.1
Mnist-10	98.3±0.2	98.5±0.1	98.8±0.2	98.3±0.1	98.6±0.3	98.7±0.2	98.6±0.1	**99.1±0.0**
Mnist-20	97.3±0.2	97.4±0.3	98.3±0.3	97.6±0.3	98.1±0.3	97.6±0.4	98.1±0.3	**98.8±0.2**
Mnist-50	95.3±0.3	94.6±0.4	96.8±0.5	96.0±0.3	97.0±0.5	94.8±1.3	97.2±0.4	**98.4±0.3**
Mnist-100	92.6±0.7	89.7±1.1	94.9±0.9	93.6±0.8	94.7±0.8	91.2±1.3	96.0±0.5	**97.3±0.6**
Fashion-1	91.4±0.2	92.3±0.2	91.7±0.4	91.4±0.2	91.7±0.3	**92.7±0.2**	91.3±0.2	92.5±0.2
Fashion-10	87.6±0.5	88.2±0.3	87.7±0.5	87.8±0.3	87.8±0.4	87.6±0.2	88.4±0.3	**89.2±0.3**
Fashion-20	85.7±0.6	85.7±0.7	85.8±0.7	85.8±0.7	85.9±0.6	85.0±0.5	86.7±0.4	**87.4±0.5**
Fashion-50	82.4±0.8	80.9±1.1	83.1±0.8	83.2±0.8	83.5±0.9	80.8±1.1	84.2±1.4	**85.7±0.6**
Fashion-100	78.8±2.1	77.3±1.0	80.1±1.6	80.5±1.2	80.5±1.2	75.6±1.5	82.2±1.6	**84.4±0.5**
SVHN-1	95.4±0.1	95.3±0.2	96.0±0.2	95.4±0.1	**96.1±0.1**	95.8±0.1	95.4±0.1	95.9±0.2
SVHN-10	88.9±0.7	87.5±0.8	91.8±0.6	90.9±0.3	92.1±0.2	90.6±0.4	91.9±0.4	**93.3±0.2**
SVHN-20	84.6±1.2	80.2±2.2	89.0±0.6	88.1±0.4	89.3±0.7	85.7±0.7	90.7±0.4	**92.0±0.5**
SVHN-50	78.0±0.5	64.9±2.4	83.4±0.5	82.4±0.6	84.5±0.7	61.3±1.8	88.4±1.4	**90.3±1.0**
SVHN-100	69.1±1.5	51.9±0.8	73.8±2.2	72.3±2.0	76.5±1.5	51.9±0.6	86.5±0.7	**87.8±0.9**
Cifar10-1	89.9±0.1	89.8±0.2	**91.2±0.2**	90.0±0.3	91.2±0.2	90.1±0.3	89.8±0.2	90.5±0.4
Cifar10-10	78.7±0.5	75.7±1.3	79.1±0.7	77.9±0.8	79.7±0.4	76.2±0.9	82.4±0.3	**82.8±0.6**
Cifar10-20	72.6±0.9	67.6±1.1	71.1±1.1	70.5±1.7	72.1±0.9	66.0±1.3	79.8±0.5	**80.4±0.6**
Cifar10-50	63.3±2.0	57.3±1.3	59.5±2.7	57.3±3.1	59.5±2.8	50.7±1.6	74.7±1.5	**77.2±0.5**
Cifar10-100	58.0±0.6	50.6±1.0	49.4±2.1	49.5±2.1	49.6±2.0	45.1±0.2	71.2±2.3	**73.6±0.9**
avgRank	6.15	6.55	3.8	5.4	3.4	6.1	3.2	1.4



(a) MNIST (b) Fashion MNIST

(c) SVHN (d) Cifar10

Fig. 1 Performance deterioration in terms of class-wise accuracy (%) with the increase of class imbalance levels. The x-axis represents different class imbalance levels, and the y-axis represents class-wise accuracy. We show class-wise accuracy between 50 and 100 to facilitate visualization.

3.3 Effect of Each Adaptive Component of ARConvL

3.3.1 Effect of Adaptive Distribution Loss

Pair-wise comparisons in terms of class-wise accuracy between ARConvL in Table 3 and the degraded ARConvL with non-adaptive β in Table 4(a) show the performance deterioration in most cases.

Table 4 Class-wise accuracy (%) of the degraded ARConvL with non-adaptive β. Each entry is the mean±std of 10 times. Better pair-wise performance compared to ARConvL in Table 3 is highlighted in bold. The last row lists average ranks (avgRank) of ARConvL vs the degraded version across datasets. Significant difference is highlighted in yellow.

	(a) Non-adaptive β			(b) Non-adaptive σ^2			(c) ARC-C
Data	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\sigma^2 = 0$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	ARC-C
Mnist-1	99.3±0.0	99.3±0.1	99.3±0.1	**99.4±0.0**	99.4±0.0	**99.4±0.1**	99.3±0.0
Mnist-10	98.7±0.1	99.1±0.0	99.0±0.1	99.0±0.1	99.0±0.1	99.1±0.1	99.1±0.1
Mnist-20	98.1±0.3	**98.9±0.1**	**98.9±0.1**	**98.9±0.1**	98.8±0.2	**98.9±0.2**	**98.9±0.1**
Mnist-50	97.1±0.4	**98.5±0.2**	**98.6±0.2**	98.4±0.2	**98.4±0.3**	**98.5±0.2**	98.4±0.2
Mnist-100	95.7±0.5	**97.6±0.3**	**97.8±0.3**	**97.5±0.4**	**97.6±0.4**	97.3±0.6	**97.5±0.5**
Fashion-1	91.7±0.2	92.5±0.2	92.3±0.2	92.1±0.1	92.4±0.2	92.4±0.2	92.0±0.2
Fashion-10	88.2±0.2	**89.4±0.3**	89.0±0.5	89.0±0.4	89.2±0.3	89.2±0.4	86.4±1.0
Fashion-20	86.3±0.6	**87.8±0.7**	87.4±0.8	**87.5±0.8**	87.3±1.1	87.4±0.6	85.0±1.4
Fashion-50	83.9±0.7	85.1±1.4	**85.7±0.7**	**86.0±0.5**	**85.9±0.5**	85.4±1.1	83.7±1.1
Fashion-100	82.2±1.5	83.8±0.8	84.0±1.1	84.1±1.3	83.9±1.4	83.8±1.2	82.8±1.4
SVHN-1	**96.3±0.1**	95.7±0.3	94.9±0.7	95.4±0.2	95.6±0.2	**95.9±0.2**	14.8±1.0
SVHN-10	93.1±0.5	93.3±0.4	92.1±1.5	92.1±0.4	92.3±0.6	92.9±0.3	55.4±34.1
SVHN-20	91.4±0.5	**92.1±0.4**	91.9±0.6	90.4±1.1	90.3±1.7	91.4±1.4	76.1±21.0
SVHN-50	88.5±1.2	89.6±1.3	90.0±1.5	88.3±1.0	89.0±1.2	89.6±0.5	80.6±1.8
SVHN-100	84.4±2.9	87.4±0.7	86.3±4.1	85.7±2.1	85.9±1.6	86.3±2.7	78.1±4.1
Cifar10-1	**92.2±0.2**	90.4±0.4	89.8±0.5	89.5±0.4	90.2±0.4	**90.5±0.4**	71.9±6.6
Cifar10-10	**83.1±0.5**	**83.4±0.6**	82.3±0.9	80.8±1.0	82.1±0.6	82.5±0.6	67.5±1.2
Cifar10-20	79.6±0.5	**80.8±0.6**	80.2±1.0	77.9±1.3	79.6±0.6	80.2±0.7	66.2±1.3
Cifar10-50	73.8±1.5	76.9±0.7	77.1±0.7	75.1±1.3	76.2±1.0	76.0±1.5	64.4±1.8
Cifar10-100	67.4±2.8	71.2±3.2	**74.0±0.5**	71.3±2.4	72.3±2.1	71.8±2.5	63.1±2.9
avgRank	1.15/1.85	1.4/1.6	1.25/1.75	1.25/1.75	1.15/1.85	1.25/1.75	1.1/1.9

Given fixed $\beta = 0$ and $\beta = 1$, Wilcoxon signed rank tests reject H0 with p-values 0.0019 and 0.019, respectively, showing significant difference in predictive performance between ARConvL and the degraded versions. Average ranks are 1.15 and 1.25 for ARConvL vs 1.85 and 1.75 for the degraded versions, respectively. This means that adaptively learning β throughout the training epochs has significantly beneficial effect on predictive performance.

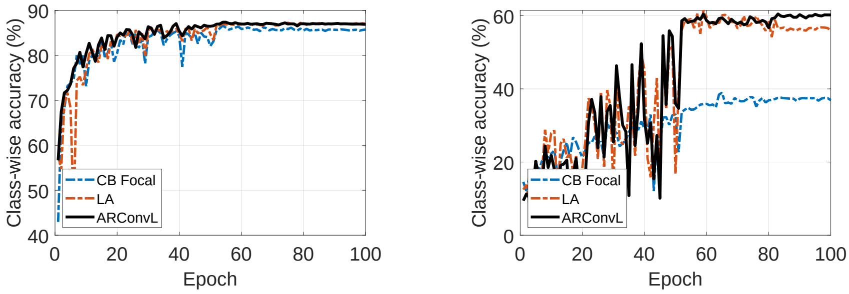
Given fixed $\beta = 0.5$, Wilcoxon signed rank test does not find significant difference between ARConvL and the degraded version with p-value 0.41. Further analyses found that on the datasets that the degraded version outperforms, performance deterioration of ARConvL is at most 0.72% in Cifar10-10; whereas on the datasets that ARConvL outperforms, performance superiority can be as high as 3.42% in Cifar10-100, with the average improvement at 0.57%. This indicates that the degraded ARConvL may cause relatively large performance decline compared to the small performance improvement it may have.

Overall, the experimental investigation shows the effectiveness of the adaptive distribution loss, in view of the adaptive β, on retaining good performance in multi-class imbalance learning.

3.3.2 Effect of Adaptive Margin

Pair-wise comparisons in terms of class-wise accuracy between ARConvL in Table 3 and the degraded ARConvL with non-adaptive σ^2 in Table 4(b) show the performance deterioration in the vast majority of cases.

Given σ^2 with those fixed values, Wilcoxon signed rank tests reject H0 with p-values 0.0022, 0.0022, and 0.0028, respectively, showing significant difference in predictive performance between ARConvL and the degraded versions with non-adaptive σ^2. Performance comparisons in terms of average ranks further show the significance of



(a) CelebA (b) iNaturalist 2018

Fig. 2 Training curves of ARConvL, LA, and CB Focal on CelebA (left) and iNaturalist 2018 (right).

such performance deterioration of the degraded ARConvL. This means that adaptively learning σ^2 throughout the training epochs has significantly beneficial effect on predictive performance, demonstrating the effectiveness of the adaptive margin on retaining good performance in multi-class imbalance learning.

3.3.3 Effect of Loss for Class Centers

Performance comparisons in terms of class-wise accuracy between ARConvL in Table 3 and the degraded ARC-C in Table 4(c) show the performance deterioration in almost all cases.

Wilcoxon signed rank test rejects H0 with p-value $3.38 \cdot 10^{-4}$, showing significant difference in predictive performance between ARConvL and the degraded ARC-C. Performance comparisons in terms of average ranks further show the significance of such performance deterioration eliminating the loss for class centers, demonstrating the effectiveness of the loss for class centers in multi-class imbalance learning.

3.4 Utility in Large-Scale Datasets

Training curves on those large-scale datasets are shown in Fig. 2. Fig. 2(a) shows that ARConvL outperforms CB Focal across all training epochs; ARConvL yields better or similar performance compared to LA and it can converge faster than LA within 52 epochs. Fig. 2(b) shows similar experimental results: ARConvL achieves better class-wise accuracy at most training epochs and possesses better convergence than its competitors. In particular, between the training epoch 52 and 78, LA and ARConvL achieve similar performance, and after the training epoch 82, ARConvL outperforms LA. Therefore, experimental results on two large-scale datasets show the utility of the proposed ARConvL over its competitors.