Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, Peter Flach

NeurIPS 2019

UNIVERSITY OF TARTU
University of BRISTOL
Universidade Federal da Paraíba ESTATÍSTICA
sphere
The Alan Turing Institute

Notion of Calibration

A probabilistic classifier p^ is:

- Multiclass-calibrated if for any prediction vector $q = (q_1, \ldots, q_k) \in \nabla_k$, proportions of classes

$$P(Y = i \mid \hat{\mathbf{p}}(X) = \mathbf{q}) = q_i \qquad for \;\; i = 1, \ldots, k.$$

- Classwise-calibrated if

$$P(Y = i \mid \hat{p}_i(X) = q_i) = q_i.$$

- Confidence-calibrated if $P\Big(Y = argmax\big(\hat{\mathbf{p}}(X)\big) \;\Big|\; max\big(\hat{\mathbf{p}}(X)\big) = c\Big) = c.$

Evaluation of Calibration

Some empirical measures of calibration are

confidence-ECE (place here equation)

$$\text{classwise}-\text{ECE} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{m}\frac{|B_{i,j}|}{n}|y_j(B_{i,j}) - \hat{p}_j(B_{i,j})|$$

Ever proper loss is minimised by the canonical calibration function

P-test versions for conf-ECE and cw-ECE

Temperature Scaling

Temperature scaling proposed by [9] learns a temperature-parameter t > 0 wich decreases (t > 1) or increases (t < 1) the confidence of a model with softmax output.

$$\sigma_{\mathrm{SM}}(\mathbf{z}_i)^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_{j=1}^{K} \exp(z_i^{(j)})}, \quad \hat{p}_i = \max_k \sigma_{\mathrm{SM}}(\mathbf{z}_i)^{(k)}$$

$$\hat{q}_i = \max_k \sigma_{\mathrm{SM}}(\mathbf{z}_i/T)^{(k)}.$$

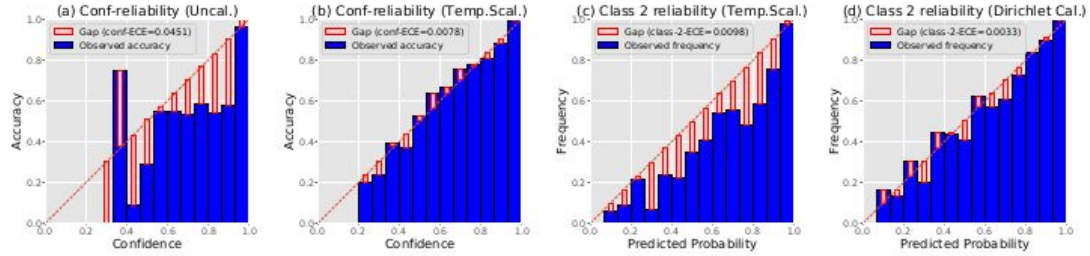


Figure 1: Reliability diagrams of `c10_resnet_wide32` on CIFAR-10: (a) confidence-reliability before calibration; (b) confidence-reliability after temperature scaling; (c) classwise-reliability for class 2 after temperature scaling; (d) classwise-reliability for class 2 after Dirichlet calibration.

UNIVERSITY OF TARTU 1632 · University of BRISTOL · Universidade Federal da Paraíba ESTATÍSTICA · sphere · The Alan Turing Institute · IPS

Dirichlet Calibration

We consider the distribution of prediction vectors p^(x) separately on instances of each class, and assume these are Dirichlet distributions with different parameters:

$$\hat{\mathbf{p}}(X) \mid Y = j \sim \mathrm{Dir}(\alpha^{(j)})$$

generative parametrisation: $\hat{\mu}_{DirGen}(\mathbf{q};\alpha,\pi) = (\pi_1 f_1(\mathbf{q}),\ldots,\pi_k f_k(\mathbf{q}))/z$

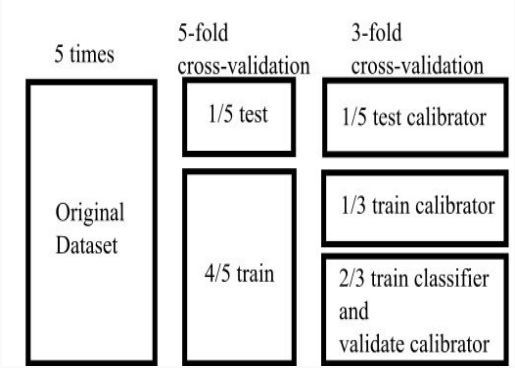
linear parametrisation: $\hat{\mu}_{DirLin}(\mathbf{q};\mathbf{W},\mathbf{b}) = \sigma(\mathbf{W}\ln\mathbf{q}+\mathbf{b})$

canonical parametrisation: $\hat{\mu}_{Dir}(\mathbf{q};\mathbf{A},\mathbf{c}) = \sigma(\mathbf{A}\ln\frac{\mathbf{q}}{1/k}+\ln\mathbf{c})$

All parameterizations are equal, i.e. they contain exactly the same calibration maps.

Experiment non-neural: Setting

- 21 datasets and 11 classifiers = 231 settings
  - Logistic, nbayes, forest, adas, lda, qda, tree, knn, mlp, svc-linear, svc-rbf
- 6 calibration methods:
  - OvR_Isotonic, OvR_Width_Bin, OvR_Freq_Bin, OvR_Beta, Temp_Scaling, Dirichlet_L2
- 8 evaluation measures
- 5 times 5-fold-crossval.
  - Inner 3-fold-crossval.



dataset	n_samples	n_features	n_classes
abalone	4177	8	3
balance-scale	625	4	3
car	1728	6	4
cleveland	297	13	5
dermatology	358	34	6
glass	214	9	6
iris	150	4	3
landsat-satellite	6435	36	6
libras-movement	360	90	15
mfeat-karhunen	2000	64	10
mfeat-morphological	2000	6	10
mfeat-zernike	2000	47	10
optdigits	5620	64	10
page-blocks	5473	10	5
pendigits	10992	16	10
segment	2310	19	7
shuttle	101500	9	7
vehicle	846	18	4
vowel	990	10	11
waveform-5000	5000	40	3
yeast	1484	8	10

Experiment non-neural: Results

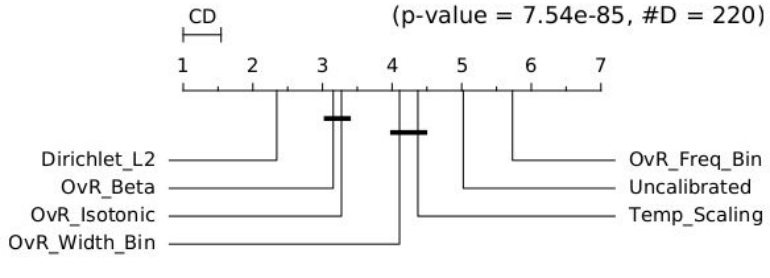
Table 1: Ranking of calibration methods for **p-cw-ECE** (Friedman's test significant with p-value $7.54e^{-85}$).

	DirL2	Beta	FreqB	Isot	WidB	TempS	**Uncal**
adas	**2.4**	3.2	4.1	4.2	3.9	5.0	5.2
forest	3.5	**2.3**	5.7	3.0	3.6	5.0	5.0
knn	2.5	4.0	4.5	**2.1**	3.2	5.8	6.0
lda	**1.9**	3.1	5.8	3.0	3.5	5.0	5.8
logistic	**2.2**	2.8	6.4	3.0	4.2	3.9	5.5
mlp	**2.2**	2.9	6.7	4.0	5.2	3.0	4.1
nbayes	**1.4**	3.6	4.8	2.6	4.2	5.3	6.1
qda	**2.2**	2.8	6.3	2.5	3.8	4.8	5.6
svc-linear	**2.3**	2.7	6.7	3.8	4.0	3.7	4.8
svc-rbf	**2.9**	3.0	6.3	3.5	4.1	3.9	4.3
tree	**2.4**	4.3	5.9	4.2	5.2	3.0	3.0
avg rank	**2.34**	3.15	5.73	3.27	4.11	4.37	5.02

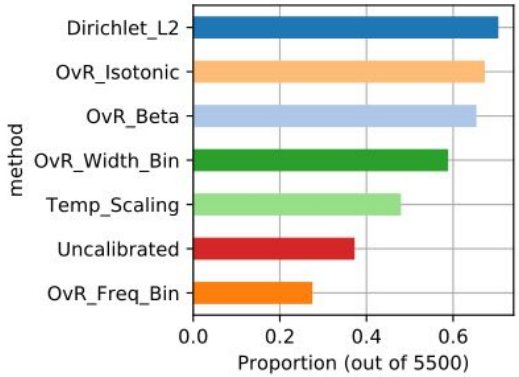
Table 2: Ranking of calibration methods for **log-loss** (p-value $4.39e^{-77}$).

DirL2	Beta	FreqB	Isot	WidB	TempS	**Uncal**
1.4	3.1	3.2	4.3	3.5	5.9	6.6
4.2	**1.9**	4.7	4.1	2.9	5.2	5.2
3.8	4.8	3.0	**1.6**	2.0	6.5	6.5
1.6	2.2	5.2	5.2	3.5	4.6	5.7
1.3	2.1	5.8	6.1	3.5	3.6	5.6
2.2	2.3	6.5	6.2	4.7	2.9	3.4
1.1	3.4	3.4	4.0	4.4	5.5	6.3
1.7	2.7	5.6	4.6	3.4	4.2	5.8
1.3	2.3	6.1	6.1	4.3	3.0	4.8
2.6	**2.2**	4.3	4.8	4.5	4.0	5.6
3.9	5.1	3.4	**2.1**	2.4	5.6	5.6
2.25	2.92	4.66	4.48	3.54	4.61	5.54

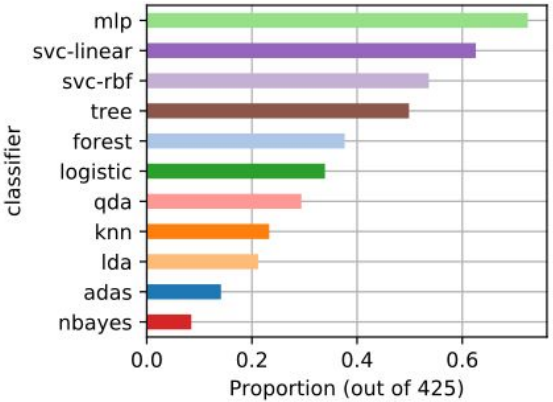
Experiment non-neural: Results



(a) p-cw-ECE critical difference



(b) p-cw-ECE for calibrators



(c) p-cw-ECE for classifiers

UNIVERSITY OF TARTU

University of BRISTOL

Universidade Federal da Paraíba ESTATÍSTICA

sphere

The Alan Turing Institute

Experiment Deep Neural Networks: Setting

- 3 datasets
  - CIFAR-10, CIFAR-100, SVHN
- 11 convolutional NNs
  - ResNet 110, ResNet 110 SD, ResNet 152 SD, WideNet 32, LeNet 5
- 3 pretrained CNNs
- 5 calibration methods:
  - TempS, Dir-L2, Dir-ODIR, VecS, MS-ODIR
- 8 evaluation measures
- 5-fold-crossval.
  - Optimal regularisation parameters

Experiment Deep Neural Networks: Results

Table 3: Scores and ranking of calibration methods for **cw-ECE**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.104_6	0.044_4	0.043_2	0.045_5	$\mathbf{0.043_1}$	0.044_3
c10_densenet40	0.114_6	0.040_5	$\mathbf{0.034_1}$	0.037_4	0.036_2	0.037_3
c10_lenet5	0.198_6	0.171_5	$\mathbf{0.052_1}$	0.059_4	0.057_2	0.059_3
c10_resnet110	0.098_6	0.043_5	$\mathbf{0.032_1}$	0.039_4	0.037_3	0.036_2
c10_resnet110_SD	0.086_6	0.031_4	0.031_5	0.029_3	0.027_2	$\mathbf{0.027_1}$
c10_resnet_wide32	0.095_6	0.048_5	0.032_3	0.029_2	0.032_4	$\mathbf{0.029_1}$
c100_convnet	0.424_6	$\mathbf{0.227_1}$	0.402_5	0.240_3	0.241_4	0.240_2
c100_densenet40	0.470_6	0.187_2	0.330_5	$\mathbf{0.186_1}$	0.189_3	0.191_4
c100_lenet5	0.473_6	0.385_5	0.219_4	0.213_2	$\mathbf{0.203_1}$	0.214_3
c100_resnet110	0.416_6	0.201_3	0.359_5	$\mathbf{0.186_1}$	0.194_2	0.203_4
c100_resnet110_SD	0.375_6	0.203_4	0.373_5	0.189_3	$\mathbf{0.170_1}$	0.186_2
c100_resnet_wide32	0.420_6	0.186_4	0.333_5	0.180_2	$\mathbf{0.171_1}$	0.180_3
SVHN_convnet	0.159_6	0.038_4	0.043_5	0.026_2	$\mathbf{0.025_1}$	0.027_3
SVHN_resnet152_SD	0.019_2	$\mathbf{0.018_1}$	0.022_6	0.020_3	0.021_5	0.021_4
Average rank	5.71	3.71	3.79	2.79	2.29	2.71

Table 4: Scores and ranking of calibration methods for **log-loss**.

		general-purpose calibrators			calibrators using logits	
	Uncal	TempS	Dir-L2	Dir-ODIR	VecS	MS-ODIR
c10_convnet	0.391_6	$\mathbf{0.195_1}$	0.197_4	0.195_2	0.197_5	0.196_3
c10_densenet40	0.428_6	0.225_5	$\mathbf{0.220_1}$	0.224_4	0.223_3	0.222_2
c10_lenet5	0.823_6	0.800_5	0.744_2	0.744_3	0.747_4	$\mathbf{0.743_1}$
c10_resnet110	0.358_6	0.209_5	$\mathbf{0.203_1}$	0.205_3	0.206_4	0.204_2
c10_resnet110_SD	0.303_6	0.178_5	0.177_4	0.176_3	0.175_2	$\mathbf{0.175_1}$
c10_resnet_wide32	0.382_6	0.191_5	0.185_4	0.182_2	0.183_3	$\mathbf{0.182_1}$
c100_convnet	1.641_6	$\mathbf{0.942_1}$	1.189_5	0.961_2	0.964_4	0.961_3
c100_densenet40	2.017_6	1.057_2	1.253_5	1.059_4	1.058_3	$\mathbf{1.051_1}$
c100_lenet5	2.784_6	2.650_5	2.595_4	2.490_2	2.516_3	$\mathbf{2.487_1}$
c100_resnet110	1.694_6	1.092_3	1.212_5	1.096_4	1.089_2	$\mathbf{1.074_1}$
c100_resnet110_SD	1.353_6	0.942_3	1.198_5	0.945_4	$\mathbf{0.923_1}$	0.927_2
c100_resnet_wide32	1.802_6	0.945_3	1.087_5	0.953_4	0.937_2	$\mathbf{0.933_1}$
SVHN_convnet	0.205_6	0.151_5	0.142_3	0.138_2	0.144_4	$\mathbf{0.138_1}$
SVHN_resnet152_SD	0.085_6	$\mathbf{0.079_1}$	0.085_5	0.080_2	0.081_4	0.081_3
Average rank	6.0	3.5	3.79	2.93	3.14	1.64

Conclusion

1. Dirichlet calibration: New parametric general-purpose multiclass calibration method
    a. Natural extension of two-class Beta calibration
    b. Easy to implement with multinomial logistic regression on log-transformed class probabilities
2. Best or tied best performance with 21 datasets x 11 classifiers
3. Advances state-of-the-art on Neural Networks by introducing ODIR regularisation

Future work

1. Which neural architectures and training methods have temperature scaling as a canonical calibration function
2. Use other distributions of the exponential family
3. Investigate scores coming from mixtures of distributions per class

Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, Peter Flach

NeurIPS 2019

UNIVERSITY OF TARTU
University of BRISTOL
Universidade Federal da Paraíba ESTATÍSTICA
sphere
The Alan Turing Institute