Supplementary Material for: "Integrating unsupervised clustering and label-specific oversampling to tackle imbalanced multi-label data"

Payel Sadhukhan[1][0000−0001−7795−3385], Arjun Pakrashi[2][0000−0002−9605−6839], Sarbani Palit[3][0000−0002−4105−6452], and Brian Mac Namee[2][0000−0003−2518−0274]

[1] TCG CREST Kolkata, India
payel0410@gmail.com
[2] School of Computer Science, University College Dublin, Ireland
{arjun.pakrashi,brian.macnamee}@ucd.ie
[3] Indian Statistical Institute Kolkata, India
palitsarbani@gmail.com

Tables 1 and 2 shows the label-based macro-average F-Score and label-based macro-averaged AUC results respectively[4], along with the relative ranks in brackets (lower ranks are better) of the algorithms compared for each dataset. The last row of both tables indicate the average rank for the algorithms. The best values are highlighted in boldface.

Also, to further analyse the differences between the algorithms, we have performed a non-parametric statistical test for a multiple classifier comparison test. We have performed a Friedman test with Finner p-value adjustments, and the critical difference plots from the test results are shown in Figure 2 and 1. The detailed results of the statistical tests a are shown in Tables 3 and 4 for the label-based macro-averaged F-Score and label-based macro-averaged AUC, respectively. The lower triangle of Tables 3 and 4 indicate the corrected p-values of the test. The upper diagonal shows a simple win/lose/tie count (algorithm on the row w.r.t. algorithm on the column). Statistical tests are summarised in the critical difference plots in Figures 1 and 2.

[4] Note that results for Table 2 does not have the results RML [**?**] as the implementation does not provide prediction scores.

Table 1: Results. Each cell indicates the averaged *label-based macro-averaged F-Scores* scores (best score in bold) along with the relative rank of the corresponding algorithm in brackets. The last row indicates the overall average ranks.

	UCLSO	COCOA	THRSEL	IRUS	SMOTE-EN	RML	BR	CLR	ECC	RAkEL
yeast	**0.505** (1)	0.461 (3)	0.427 (5)	0.426 (6)	0.436 (4)	0.471 (2)	0.409 (9)	0.413 (8)	0.389 (10)	0.420 (7)
emotions	0.658 (2)	**0.666** (1)	0.560 (9)	0.622 (5)	0.575 (8)	0.645 (3)	0.550 (10)	0.595 (7)	0.638 (4)	0.613 (6)
medical	**0.783** (1)	0.759 (2)	0.733 (3.5)	0.537 (10)	0.700 (8)	0.707 (7)	0.718 (6)	0.724 (5)	0.733 (3.5)	0.672 (9)
cal500	0.273 (2)	0.210 (5)	0.252 (3)	**0.277** (1)	0.235 (4)	0.209 (6)	0.169 (8)	0.081 (10)	0.092 (9)	0.193 (7)
rcv1-s1	**0.443** (1)	0.364 (3)	0.292 (5)	0.252 (8)	0.313 (4)	0.387 (2)	0.285 (6)	0.227 (9)	0.192 (10)	0.272 (7)
rcv1-s2	**0.432** (1)	0.342 (3)	0.275 (5)	0.234 (8)	0.305 (4)	0.363 (2)	0.272 (6)	0.226 (9)	0.173 (10)	0.263 (7)
rcv1-s3	**0.480** (1)	0.339 (3)	0.275 (5)	0.225 (8)	0.302 (4)	0.371 (2)	0.271 (6)	0.211 (9)	0.163 (10)	0.257 (7)
enron	**0.352** (1)	0.342 (2)	0.291 (5)	0.293 (4)	0.266 (8)	0.307 (3)	0.246 (9)	0.244 (10)	0.268 (6)	0.267 (7)
bibtex	**0.442** (1)	0.318 (3)	0.303 (4)	0.253 (8)	0.283 (5)	0.326 (2)	0.263 (7)	0.265 (6)	0.212 (10)	0.252 (9)
llog	**0.181** (1)	0.082 (6)	0.096 (3)	0.124 (2)	0.095 (4.5)	0.095 (4.5)	0.031 (7)	0.024 (8)	0.022 (10)	0.023 (9)
corel5k	0.209 (2)	0.196 (3)	0.146 (4)	0.105 (6)	0.125 (5)	**0.215** (1)	0.089 (7)	0.049 (10)	0.054 (9)	0.084 (8)
slashdot	**0.443** (1)	0.374 (2)	0.355 (4)	0.257 (10)	0.366 (3)	0.343 (5)	0.291 (8)	0.290 (9)	0.304 (6)	0.296 (7)
Avg. rank	**1.25**	3.00	4.62	6.33	5.12	3.29	7.42	8.33	8.12	7.5

Table 2: Results. Each cell indicates the averaged *Label-based macro-averaged AUC* scores (best score in bold) along with the relative rank of the corresponding algorithm in brackets. The last row indicates average ranks.

	UCLSO	COCOA	THRSEL	IRUS	SMOTE-EN	BR	CLR	ECC	RAkEL
yeast	0.666 (3)	**0.711** (1)	0.576 (8.5)	0.658 (4)	0.582 (7)	0.576 (8.5)	0.650 (5)	0.705 (2)	0.641 (6)
emotions	0.819 (3)	0.844 (2)	0.687 (8.5)	0.802 (4)	0.698 (7)	0.687 (8.5)	0.796 (6)	**0.850** (1)	0.797 (5)
medical	**0.967** (1)	0.964 (2)	0.869 (7.5)	0.955 (3.5)	0.873 (6)	0.869 (7.5)	0.955 (3.5)	0.952 (5)	0.856 (9)
cal500	0.550 (4)	0.558 (2)	0.509 (8.5)	0.545 (5)	0.512 (7)	0.509 (8.5)	**0.561** (1)	0.557 (3)	0.528 (6)
rcv1-s1	**0.919** (1)	0.889 (3)	0.643 (7.5)	0.882 (4)	0.626 (9)	0.643 (7.5)	0.891 (2)	0.881 (5)	0.728 (6)
rcv1-s2	**0.912** (1)	0.882 (2.5)	0.640 (7.5)	0.880 (4)	0.622 (9)	0.640 (7.5)	0.882 (2.5)	0.874 (5)	0.721 (6)
rcv1-s3	**0.956** (1)	0.880 (2)	0.633 (7.5)	0.872 (4.5)	0.628 (9)	0.633 (7.5)	0.877 (3)	0.872 (4.5)	0.718 (6)
enron	0.719 (5)	**0.752** (1)	0.597 (8.5)	0.738 (3)	0.619 (7)	0.597 (8.5)	0.720 (4)	0.750 (2)	0.650 (6)
bibtex	0.844 (4)	0.877 (2)	0.673 (8.5)	**0.894** (1)	0.706 (6)	0.673 (8.5)	0.811 (5)	0.873 (3)	0.696 (7)
llog	**0.721** (1)	0.663 (4)	0.518 (7.5)	0.676 (2)	0.561 (6)	0.518 (7.5)	0.612 (5)	0.673 (3)	0.514 (9)
corel5k	0.695 (4)	0.718 (3)	0.559 (7.5)	0.687 (5)	0.596 (6)	0.559 (7.5)	**0.740** (1)	0.723 (2)	0.552 (9)
slashdot	**0.806** (1)	0.774 (2)	0.632 (8.5)	0.753 (4)	0.714 (6)	0.632 (8.5)	0.742 (5)	0.765 (3)	0.638 (7)
Avg. ranks	2.42	**2.21**	8.00	3.67	7.08	8.00	3.58	3.21	6.83

Table 3: Statistical analysis and comparison of *label-based macro-averaged F-Scores*. Upper diagonal: win/lose/tie. Lower diagonal: Results of the Friedman rank test with Finner p-value correction. * $\alpha = 0.1$, ** $\alpha = 0.05$ and *** $\alpha = 0.01$

	UCLSO	COCOA	THRSEL	IRUS	SMOTE.EN	RML	BR	CLR	ECC	RAkEL
UCLSO		11/1/0	12/0/0	11/1/0	12/0/0	11/1/0	12/0/0	12/0/0	12/0/0	12/0/0
COCOA	0.219		10/2/0	10/2/0	10/2/0	5/7/0	12/0/0	12/0/0	12/0/0	12/0/0
THRSEL	0.017 **	0.248		8/4/0	6/6/0	4/8/0	12/0/0	11/1/0	10/1/1	11/1/0
IRUS	0.000 ***	0.017 **	0.227		4/8/0	2/10/0	6/6/0	9/3/0	9/3/0	7/5/0
SMOTE.EN	0.006 ***	0.143	0.711	0.400		2/9/1	11/1/0	10/2/0	9/3/0	10/2/0
RML	0.159	0.827	0.353	0.031 **	0.206		11/1/0	11/1/0	11/1/0	12/0/0
BR	0.000 ***	0.001 ***	0.046 **	0.442	0.112	0.003 ***		8/4/0	8/4/0	7/5/0
CLR	0.000 ***	0.000 ***	0.008 ***	0.164	0.022 **	0.000 ***	0.516		6/6/0	3/9/0
ECC	0.000 ***	0.000 ***	0.013 **	0.212	0.032 **	0.000 ***	0.610	0.872		4/8/0
RAkEL	0.000 ***	0.001 ***	0.041 **	0.411	0.100	0.002 ***	0.946	0.551	0.647	

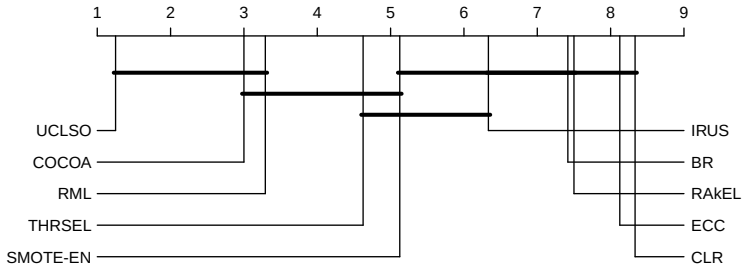

Fig. 1: Critical difference plot of Friedman rank sum test with Finner p-value correction on *label-basesd macro-averaged F-Scores*. The scale indicates the average ranks. The methods which are not connected with the horizontal lines are significantly different with a significance level of $\alpha = 0.05$.

Table 4: Statistical analysis and comparison of *label-based macro-averaged AUC*. Upper diagonal: win/lose/tie. Lower diagonal: Results of the Friedman rank test with Finner p-value correction. * $\alpha = 0.1$, ** $\alpha = 0.05$ and *** $\alpha = 0.01$

	UCLSO	COCOA	THRSEL	IRUS	SMOTE.EN	BR	CLR	ECC	RAkEL
UCLSO		6/6/0	12/0/0	10/2/0	12/0/0	12/0/0	9/3/0	6/6/0	12/0/0
COCOA	0.868		12/0/0	10/2/0	12/0/0	12/0/0	8/3/1	9/3/0	12/0/0
THRSEL	0.000 ***	0.000 ***		0/12/0	3/9/0	0/0/12	0/12/0	0/12/0	3/9/0
IRUS	0.380	0.306	0.000 ***		12/0/0	12/0/0	6/5/1	5/6/1	12/0/0
SMOTE.EN	0.000 ***	0.000 ***	0.495	0.004 ***		9/3/0	0/12/0	0/12/0	5/7/0
BR	0.000 ***	0.000 ***	1.000	0.000 ***	0.495		0/12/0	0/12/0	3/9/0
CLR	0.410	0.332	0.000 ***	0.945	0.004 ***	0.000 ***		6/6/0	11/1/0
ECC	0.543	0.461	0.000 ***	0.736	0.001 ***	0.000 ***	0.778		12/0/0
RAkEL	0.000 ***	0.000 ***	0.410	0.008 ***	0.849	0.410	0.007 ***	0.003 ***	

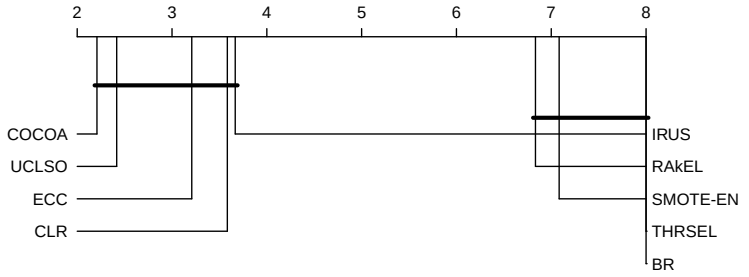

Fig. 2: Critical difference plot of Friedman rank sum test with Finner p-value correction on *label-basesd macro-averaged AUC* scores. The scale indicates the average ranks. The methods which are not connected with the horizontal lines are significantly different with a significance level of $\alpha = 0.05$.