Supplemental Material: Interactive Reweighting for Biased Training Samples

Appendix A: Effects of Different Threshold ε for Discretizing the Influence Values into Three Categories: Positive, Neutral, and Negative

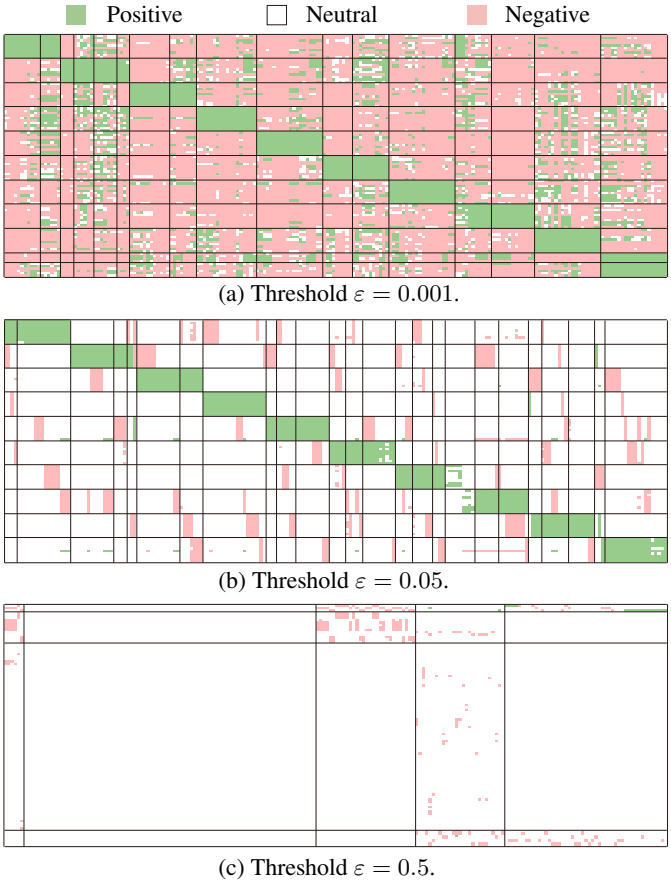
Before the co-clustering, we need to discretize the continuous influence values into three categories: positive, neutral, and negative. We experiment with the four datasets used in Sec. 6 to decide the best threshold ε for the positive ($\geq \varepsilon$), neutral (between $-\varepsilon$ and ε), and negative categories ($\leq -\varepsilon$). To evaluate the quality of discretization results with different thresholds, we first generate the ground-truth categories of the influence values. The ground-truth categories are obtained based on the labels of the images. For example, if both the validation samples and training samples are clean and of the same labels, the ground-truth categories of the influence values between them are positive. With the ground-truth categories, we evaluate the quality of the discretization results using the macro F1-score, which is the average F1-scores of all categories. We choose the macro F1-score because it is suitable for unbalanced categories distribution [1], which applies to the influence values, as most of them are of neural category. As shown in Table S1, $\varepsilon = 0.05$ works the best in all the datasets.

Table S1: Macro F1-score comparison using different thresholds.

Dataset	Threshold (ε)					
	0.001	0.005	0.01	0.05	0.1	0.5
CIFAR10	0.434	0.719	0.845	**0.862**	0.631	0.306
CIFAR100	0.365	0.539	0.646	**0.832**	0.821	0.490
Clothing	0.349	0.564	0.654	**0.779**	0.698	0.320
OCT	0.501	0.663	0.724	**0.732**	0.553	0.254

We also examined the co-clustering results using three different thresholds: the smallest threshold (0.001), the appropriate threshold (0.05), and the largest threshold (0.5). Fig. S1 shows the co-clustering results on the CIFAR10 dataset. It can be seen from this figure that the numbers of validation sample clusters are 11 ($\varepsilon = 0.001$), 10 ($\varepsilon = 0.05$), and 4 ($\varepsilon = 0.5$), respectively. When $\varepsilon = 0.05$, the 10 validation sample clusters align well with the 10 classes of the CIFAR10 dataset. In comparison, when $\varepsilon = 0.001$, one class is divided into two clusters, and the remaining nine classes form individual clusters. However, when $\varepsilon = 0.5$, the co-clustering algorithm fails to generate informative clusters because nearly all influence values are discretized into neutral (Fig. S1(c)).

To understand why the co-clustering result for $\varepsilon = 0.05$ is better than that of $\varepsilon = 0.001$, we examine the associated training sample clusters. It is observed that the number of training sample clusters with $\varepsilon = 0.001$ (17) is smaller than the one with $\varepsilon = 0.05$ (25). This is because most of the influence values are discretized into negative when $\varepsilon = 0.001$ (Fig. S1(a)), which hides the subtle differences between training samples. These differences are discernable with an appropriate threshold, as highlighted by the red blocks in Fig. S1(b).



(a) Threshold $\varepsilon = 0.001$.

(b) Threshold $\varepsilon = 0.05$.

(c) Threshold $\varepsilon = 0.5$.

Fig. S1: Co-clustering results using different threshold values. Rows represent validation samples and their clusters, columns represent training samples and their clusters, and cells represent corresponding influence values.

Appendix B: Effects of Different γ in the Validation Sample Weight Adjustment

In Reweighter, model developers can adjust the weights of validation samples by indicating the desired direction of weight changes, which is formulated as the inequality constraint of the optimization problem. In our implementation, the inequality constraint is set as $w_i^v \geq (1+\gamma)\tilde{w}_i^v$ when increasing the weight and as $w_i^v \leq (1-\gamma)\tilde{w}_i^v$ when decreasing the weight, where $\tilde{w}_i^v$ is the previous weight. Table S2 shows the percentages of correctly reweighted samples using different γ values. We found that all the percentages were in the range between 0.773 and 0.776, indicating that our method is robust to the selection of γ. However, when $\gamma > 0.1$, the percentage of correctly reweighted samples slowly decreases with the increase of γ. This suggests the inappropriateness of a large γ value. In contrast, a small γ does not reflect the user intention to change the weights. Therefore, we set $\gamma = 0.1$ in our implementation.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Table S2: Percentages of the correctly reweighted samples with different γ values.

γ	0	1e-4	1e-3	1e-2	1e-1	2e-1	4e-1	8e-1
Percentage	0.775	0.775	**0.776**	0.775	**0.776**	0.775	0.774	0.773

Table S3: Accuracy comparison under the combined noise.

Dataset	# labels per class	Imbal. factor	Uniform	FSR	Ours
CIFAR10	10	5	55.1%	57.4%	**58.0%**
	10	10	54.2%	56.3%	**56.9%**
	10	20	53.0%	54.8%	**55.8%**
	10	50	50.4%	53.8%	**55.0%**
	20	5	62.4%	62.8%	**63.2%**
	20	10	61.2%	62.1%	**62.7%**
	20	20	59.5%	61.1%	**62.2%**
	20	50	58.0%	59.8%	**60.5%**
	50	5	66.9%	66.7%	**67.0%**
	50	10	65.5%	66.1%	**66.6%**
	50	20	64.4%	65.6%	**66.0%**
	50	50	62.5%	63.2%	**63.7%**
	100	5	69.8%	70.4%	**70.6%**
	100	10	68.8%	69.3%	**69.6%**
	100	20	67.7%	68.0%	**68.7%**
	100	50	65.0%	66.4%	**67.0%**
CIFAR100	10	5	29.5%	30.4%	**30.8%**
	10	10	28.2%	29.0%	**29.7%**
	10	20	27.1%	28.0%	**28.6%**
	10	50	26.2%	26.9%	**27.7%**
	20	5	34.7%	35.0%	**35.2%**
	20	10	32.9%	33.4%	**33.8%**
	20	20	31.1%	32.1%	**32.7%**
	20	50	28.8%	31.0%	**31.6%**
	50	5	41.2%	41.9%	**42.6%**
	50	10	39.3%	40.3%	**41.1%**
	50	20	36.9%	38.3%	**39.4%**
	50	50	34.7%	36.0%	**36.9%**
	100	5	48.1%	48.1%	**48.6%**
	100	10	44.8%	44.9%	**45.4%**
	100	20	41.8%	42.2%	**42.9%**
	100	50	38.3%	40.1%	**41.0%**
Clothing	N/A	5	55.9%	59.8%	**62.1%**
	N/A	10	52.6%	57.4%	**59.8%**
	N/A	20	50.1%	54.2%	**57.1%**
	N/A	50	47.9%	51.1%	**54.3%**
OCT	10	N/A	49.5%	57.4%	**59.6%**
	20	N/A	53.0%	63.8%	**65.2%**
	50	N/A	66.0%	71.3%	**73.1%**
	100	N/A	69.2%	77.0%	**78.9%**

APPENDIX C: PERFORMANCE COMPARISON IN THE COMBINED SCENARIO.

Table S3 shows the full results of the performance comparison in the combined noise scenario. It demonstrates the capability of our method in handling both noisy labels and imbalanced class distributions. For the CIFAR10 and CIFAR100 datasets, the performance gain increase with the increasing imbalanced factor and decreasing number of labeled samples per class. This is because when the label noise is more severe, and the data is more imbalanced, it is more important to guarantee the quality of validation samples. For the Clothing dataset, the performance gain does not further increase when the imbalanced factor is larger than 10. After analyzing the validation samples, it turns out that the performance bottleneck is the label noise in close-related categories, such as "sweater" and "knitwear." In that case, our method does not bring more gains in a more imbalanced dataset. For the retinal OCT data, our method also achieves a greater performance gain with the decreasing number of labeled samples per class, which is similar to the CIFAR datasets.

REFERENCES

[1] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, and M. Linden. A comparative analysis of hybrid deep learning models for human activity recognition. *Sensors*, 20(19):5707, 2020. doi: 10.3390/s20195707 1