universität wien

ECML PKDD 2023

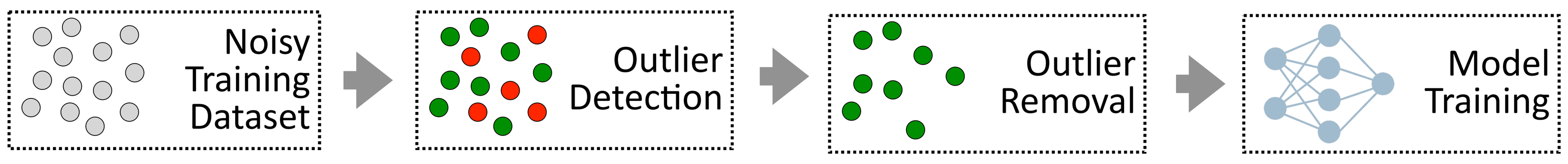
W|W|T|F

Learning with Noisy Labels by Adaptive GRAdient-Based Outlier Removal

Anastasiia Sedova*, Lena Zellinger*, Benjamin Roth
{anastasiia.sedova, lena.zellinger, benjamin.roth}@univie.ac.at
University of Vienna

Background & Motivation

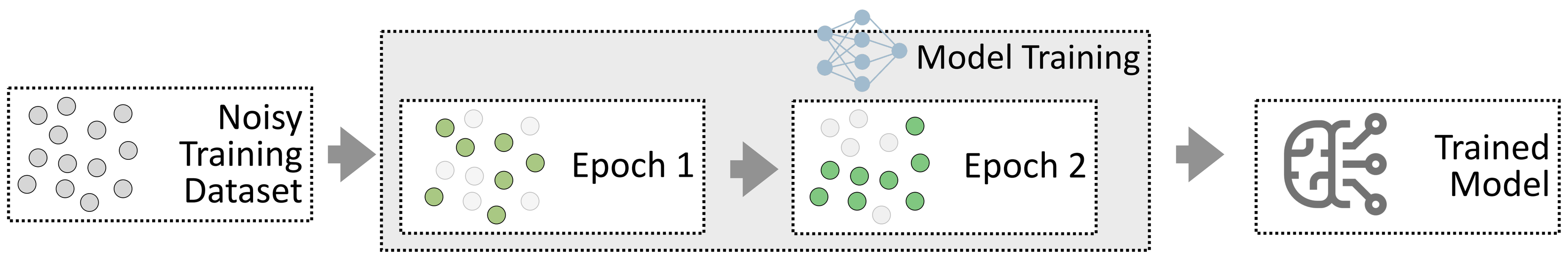
✦ Usual outlier detection is *static*: the outliers are detected before the model training.



✦ However, even the mislabeled samples can be *useful and beneficial* for the model in some training stages.

Example: *"The movie was by no means great."*– POSITIVE
This (mislabeled) sample can help a model on the early training stages to learn a useful assosication between word *great* and class POSITIVE.

✦ Instead of *static* removal of samples **before** training, we suggest to *dynamically* adjust the training set **during** training.



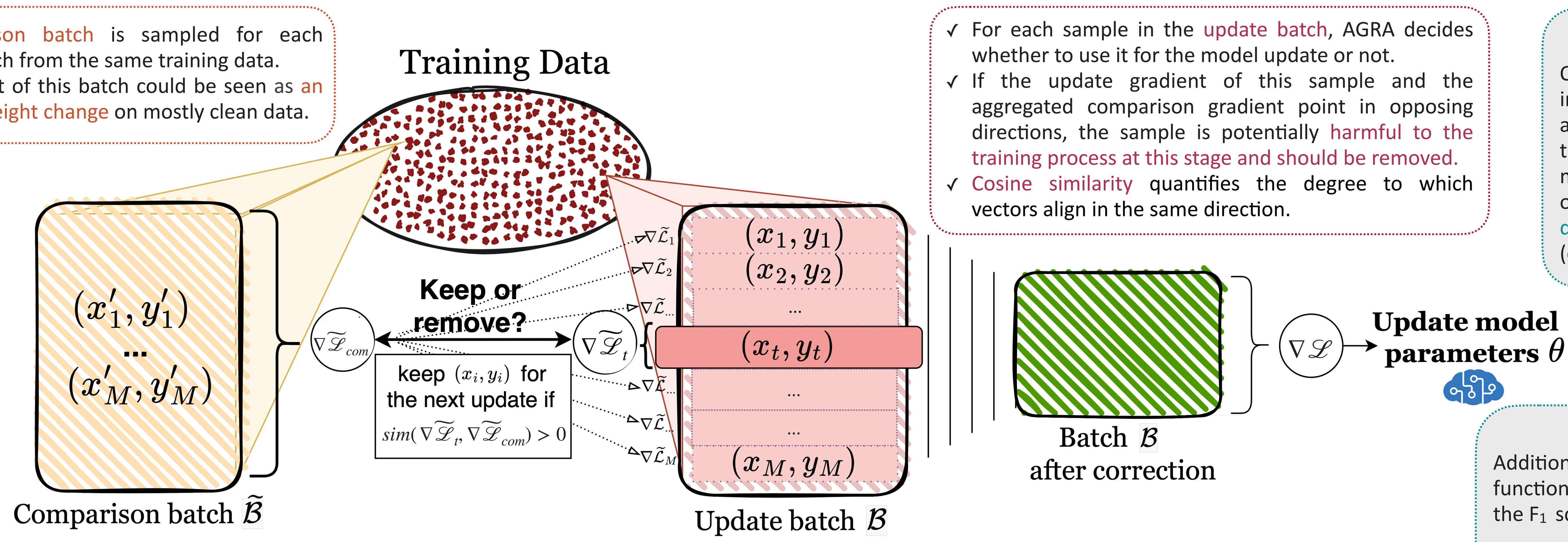
✦ Our method AGRA for Adaptive GRAdient-Based Outlier Removal decides for each sample whether it is **useful or not** for a model on the current training stage and either keeps it or removes.

AGRA Methodology

Detect the instances that would harm the model in the current training stage and filter them out before the update.

- ✓ A comparison batch is sampled for each training batch from the same training data.
- ✓ The gradient of this batch could be seen as an expected weight change on mostly clean data.

- ✓ For each sample in the update batch, AGRA decides whether to use it for the model update or not.
- ✓ If the update gradient of this sample and the aggregated comparison gradient point in opposing directions, the sample is potentially harmful to the training process at this stage and should be removed.
- ✓ Cosine similarity quantifies the degree to which vectors align in the same direction.



$$sim_{y_t} = sim\left(\nabla\widetilde{\mathcal{L}}(x_t, y_t), \nabla\widetilde{\mathcal{L}}_{com}\right) = \frac{\nabla\widetilde{\mathcal{L}}(x_t, y_t) \cdot \nabla\widetilde{\mathcal{L}}_{com}}{||\nabla\widetilde{\mathcal{L}}(x_t, y_t)||_2 \, ||\nabla\widetilde{\mathcal{L}}_{com}||_2}$$

To ensure a well-balanced comparison batch, AGRA provides class-weighted sampling. The weight for a class is computed as the inverted number of occurrences of this class in the training set.

$$\frac{1}{\sum_{t=1}^{|\mathcal{X}|} \mathbb{1}(y_t = c_k)}$$

Alternative Label

Optionally, AGRA may include alternative label as one of the options. In this case, the sample might be kept, removed or assigned to a pre-defined alternate class (e.g., the negative class).

The gradient of the corrected updated batch is used for the model update in the current training stage.

F_1 Loss Function

Additionally, we introduce F_1 loss function which aims to maximize the F_1 score.

$$\mathcal{L}_{F_{1_M}}(\mathcal{B}) = 1 - \frac{1}{K}\sum_{k=1}^{K} \frac{2\widehat{tp}_k}{2\widehat{tp}_k + \widehat{fp}_k + \widehat{fn}_k + \epsilon}$$

$$\widehat{tp}_k = \sum_{t=1}^{M} \hat{y}_{t,k} \times \mathbb{1}(y_t = c_k)$$

$$\widehat{fp}_k = \sum_{t=1}^{M} \hat{y}_{t,k} \times (1 - \mathbb{1}(y_t = c_k))$$

$$\widehat{fn}_k = \sum_{t=1}^{M} (1 - \hat{y}_{t,k}) \times \mathbb{1}(y_t = c_k)$$

NB! F_1 loss function is not mandatory; AGRA is compatible with any loss function.

Experiments & Discussion

✦ Datasets:
- 5 weakly annotated text datasets (spam detection, question classification, topic classification in low-resource languages)
- 2 image datasets (CIFAR with added noise & weakly annotated CheXpert).

✦ Baselines: 3 weakly supervised methods, 2 noisy learning methods.

✦ Logistic regression classifier with tf-idf representations.

✦ Main result: Ours outperforms all the baselines on five datasets and is the best on average on text data.

	YouTube (Acc)	SMS (F1)	TREC (Acc)	Yorùbá (F1)	Hausa (F1)	Avg.	CIFAR (Acc)	CXT (AUR)
Gold	94.8±0.8	95.4±1.0	89.5±0.3	57.3±0.4	78.5±0.3	83.1	83.6±0.0	–
No Denoising	87.4±2.7	71.7±1.4	58.7±0.5	44.6±0.4	39.7±0.8	60.4	82.4±0.2	82.7±0.1
Weak Supervision								
DP [35]	90.8±1.0	44.1±6.7	54.3±0.5	**47.8**±1.7	40.9±0.6	55.6	–	–
MeTaL [34]	92.0±0.8	18.3±7.8	50.4±1.7	38.9±3.1	45.5±1.1	49.0	–	–
FS [14]	84.8±1.2	16.3±6.0	27.2±0.1	31.9±0.7	37.6±1.0	39.6	–	–
Noisy Learning								
CORES^2 [10]	88.8±3.6	85.8±1.8	61.8±0.5	43.0±0.7	**51.2**±0.5	66.1	83.4±0.1	–
Cleanlab [32]	91.3±1.2	80.6±0.3	60.9±0.4	43.8±1.3	40.3±0.3	63.4	83.3±0.0	81.5±0.4
AGRA	**93.9**±0.7	**87.7**±1.2	**63.6**±0.7	46.9±1.5	46.2±1.6	**67.7**	**83.6**±0.0	**83.9**±0.3

✦ Ablation study: Ours outperforms the baselines in most settings.

✦ F_1-based comparison loss function is beneficial for all datasets.

✦ Weighted comparison batch sampling is especially helpful for imbalanced datasets (e.g., Hausa and TREC)

	No Weighted Sampling		Weighted Sampling	
	CE/CE	CE/F_1	CE/CE	CE/F_1
YouTube	92.0 ± 1.0	$\mathbf{93.9 \pm 0.7}$	91.9 ± 0.5	93.4 ± 0.8
YouTube†	90.5 ± 1.0	–	92.0 ± 0.7	–
SMS	79.0 ± 3.2	61.1 ± 5.2	$\mathbf{87.7 \pm 1.2}$	49.1 ± 3.0
SMS†	71.1 ± 3.1	–	86.3 ± 1.2	–
TREC	61.6 ± 0.6	62.1 ± 0.4	62.8 ± 1.1	$\mathbf{63.6 \pm 0.7}$
Yorùbá	44.3 ± 2.5	44.2 ± 1.4	43.5 ± 1.0	$\mathbf{46.9 \pm 1.5}$
Hausa	41.2 ± 0.4	40.9 ± 0.6	43.8 ± 2.8	$\mathbf{46.2 \pm 1.6}$
CheXpert	82.6 ± 0.6	$\mathbf{83.9 \pm 0.3}$	–	–
CIFAR	82.2 ± 0.2	83.5 ± 0.0	83.1 ± 0.0	$\mathbf{83.6 \pm 0.0}$

Table 3: AGRA experimental test results with different settings: use of class-weighted sampling, [training loss]/[comparison loss]. The results marked with † are obtained by AGRA with an alternative label. All results are averaged across 5 runs and reported with standard deviation.

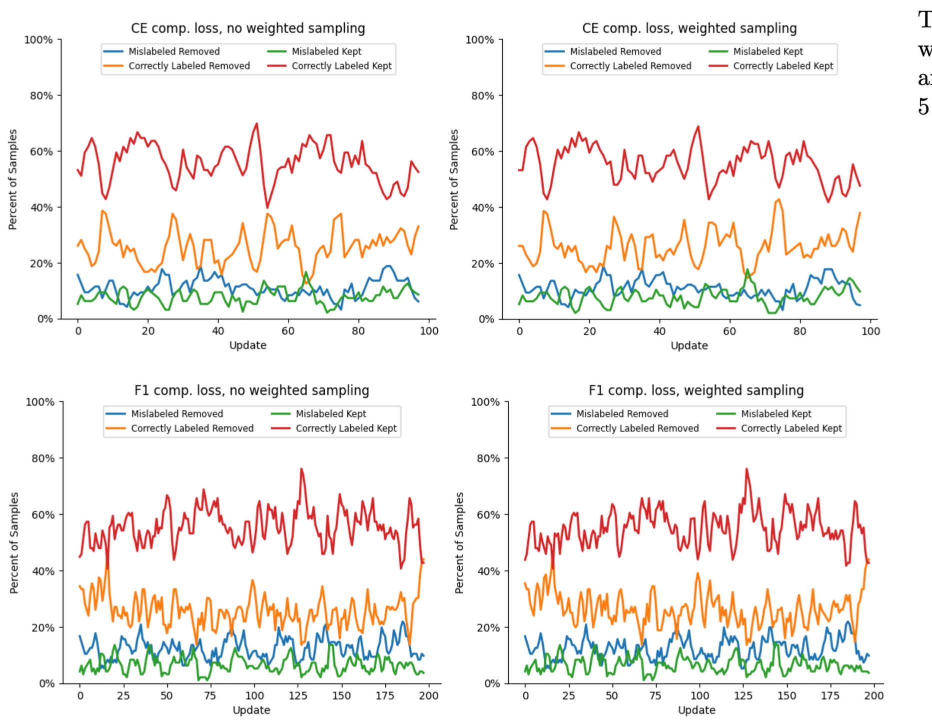


Fig. 2: Case study on the YouTube dataset. The plots represent the percentage of samples in each batch that were correctly kept, correctly removed, falsely kept and falsely removed during the training of the best-performing models for all combinations of comparison losses and sampling strategies.

✦ Case study: YouTube dataset.

✦ Notably, the amount of "falsely" kept and "falsly" removed vary greatly and even exceeds the amount of "correctly" kept and removed in some training stages.

Our main observation: correctness of removed samples appears to be not crucial for training a reliable model.