Which Is More Effective in Label Noise Cleaning, Correction or Filtering?

Gaoxia Jiang[1], Jia Zhang[1], Xuefei Bai[1], Wenjian Wang[1]*, Deyu Meng[2]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2]School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China
jianggaoxia@sxu.edu.cn, 774924867@qq.com, {baixuefei, wjwang}@sxu.edu.cn, dymeng@mail.xjtu.edu.cn

Abstract

Most noise cleaning methods adopt one of the correction and filtering modes to build robust models. However, their effectiveness, applicability, and hyper-parameter insensitivity have not been carefully studied. We compare the two cleaning modes via a rebuilt error bound in noisy environments. At the dataset level, Theorem 5 implies that correction is more effective than filtering when the cleaned datasets have close noise rates. At the sample level, Theorem 6 indicates that confident label noises (large noise probabilities) are more suitable to be corrected, and unconfident noises (medium noise probabilities) should be filtered. Besides, an imperfect hyper-parameter may have fewer negative impacts on filtering than correction. Unlike existing methods with a single cleaning mode, the proposed Fusion cleaning framework of Correction and Filtering (FCF) combines the advantages of different modes to deal with diverse suspicious labels. Experimental results demonstrate that our FCF method can achieve state-of-the-art performance on benchmark datasets.

Introduction

Large-scale datasets with sufficient and accurate labels make supervised learning models, especially for deep neural networks, possible to update their parameters and achieve excellent performances. However, these labels may be noisy due to insufficient information or inexperienced labeling workers (Han et al. 2019; Jiang et al. 2021). The label noise usually misguides the model training and raises the generalization risk (Frenay and Verleysen 2014). Cleaning noisy data and learning with label noise are critical and challenging in supervised learning and complicated real applications (Wu et al. 2020; Shu, Yuan, and Meng 2023).

Noise-robust modeling and noisy data cleaning are the main directions in dealing with label noise. Compared with noise-robust modeling, noisy data cleaning is more generic and flexible. It has been considered the most competitive noisy-label learning strategy (Kim et al. 2021; Li, Socher, and Hoi 2020). The noisy labels can be cleaned by correction or filtering. The former aims to change the corrupted labels to the true ones (Zheng et al. 2020; Wu et al. 2021; Li and Sun 2022; Li et al. 2023). And noise filtering refers to removing suspicious noisy samples (Song, Kim, and Lee 2019; Wu et al. 2020; Kim et al. 2021; Wei et al. 2022).

It has been verified that both correction and filtering are effective in dealing with label noises, but they still have respective drawbacks, such as wrong corrections and over-cleaning. Specifically, empirical results show that there are still about 15% wrong labels after correction on CIFAR-100 with 40% uniform noise (Wu et al. 2021). Although the label purity (correct rate) of the filtered dataset approaches 100%, at least 10% of correctly labeled samples are removed in filtering on CIFAR-100 with 30% pair flip noise (Wu et al. 2020). Most noise cleaning methods adopt one of the correction and filtering modes to generate robust models. How to select a proper cleaning mode is a fundamental problem in label noise cleaning.

To the best of our knowledge, correction and filtering have not been systematically compared and integrated from the error-bound perspective. We investigate three key properties: (1) *Effectiveness*. Which could yield a lower generalization error (bound)? (2) *Applicability*. What kind of noise is the cleaning mode suitable for handling? (3) *Hyper-parameter insensitivity*. Which is less sensitive to an imperfect hyper-parameter? The motivation is to guide the selection of cleaning modes for designing effective algorithms.

Error bound is a remarkable theoretical tool in machine learning. However, existing bounds are mainly deduced in noise-free conditions. We rebuild the error bound in learning with uniform label noise to compare correction and filtering. Based on this, we develop the first fusion cleaning framework of correction and filtering (FCF) to deal with different types of suspicious labels. The main contributions are:

- **Rebuilt error bound for label noise**. We rebuild the error bound for uniform label noise (Theorem 1) which can be viewed as a generalized form of the conventional bound. It provides a powerful tool to analyze the performance in learning with label noise.
- **Cleaning mode comparison and selection**. Correction and filtering are compared in effectiveness, applicability, and hyper-parameter insensitivity. It tells us what kind of noise is suitable for correction/filtering.
- **Fusion of correction and filtering.** We develop a feasible fusion cleaning framework as shown in Figure 1 by combining the advantages of correction and filtering.

*Corresponding author.

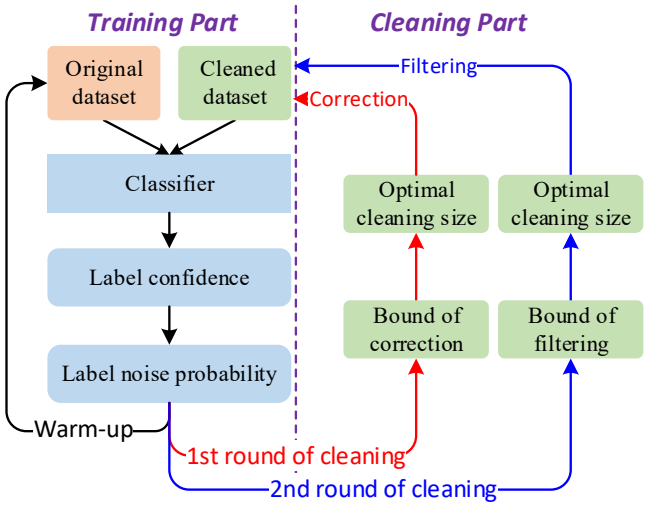


Figure 1: The proposed FCF cleaning framework. In the early stage, the classifier is trained on the original dataset to produce label confidence and estimate noise probability. Then a certain amount (k_1) of large-noise-probability (low-confidence) labels are corrected in the first round of cleaning. The corrected labels usually have median or small noise probabilities. In the second round of cleaning, a certain amount (k_2) of samples with unconfident labels (maybe wrongly corrected) will be removed from the corrected dataset. The data (cleaning part) and model (training part) can be interactively improved in the FCF framework.

Related Works

Noise-robust modeling. Noise-robust models are usually built through robust loss functions, regularization, sample weighting, robust architecture, or ensemble methods (Zhang et al. 2017; Song et al. 2020; Xia et al. 2020; Shu et al. 2019; Song et al. 2022), but their performance is far from state-of-the-art when the datasets suffer from severe label noises (Han et al. 2018; Kim et al. 2021).

Label noise correction. Many label correctors have been proposed to identify and correct noisy labels based on model predictions or feature representations. (Huang, Zhang, and Zhang 2020) predicted the correct label by tracking all historical model predictions using an exponential moving average scheme. (Zheng et al. 2020) identified a clean label according to the likelihood ratio, which was defined based on the noisy label prediction and best label prediction. (Sharma et al. 2020) detected and corrected noisy labels by utilizing the cluster relationships between all noisy samples. (Wu et al. 2020) used the geometry and topology of instance representations to aggressively collect clean samples. (Wu et al. 2021) proposed a meta-learning model, aiming at training an automatic scheme that can estimate soft labels under the guidance of a small amount of noise-free metadata. (Li and Sun 2022) designed a label correction method simultaneously updating model parameters and correcting noisy labels. Although the label quality is improved by replacing the noisy labels with the underlying true ones, some of these methods require prior knowledge about the noise rate, or they tend to be sensitive to hyper-parameters like the confidence threshold.

Label noise filtering. Label noise filtering is also known as the high-quality sample/instance selection. Conventional filters mainly contain ensemble-based and nearest-neighbor-based methods. Multiple model predictions are considered in the ensemble-based filter to obtain more accurate noise recognition (Garcia et al. 2016). These predictions can be generated by various classifiers or prediction schemes such as cross-validation (Sáez et al. 2016; Northcutt, Athalye, and Mueller 2021). (Northcutt, Jiang, and Chuang 2021) utilized confidence learning to detect noisy labels on some commonly used datasets. Noisy labels are filtered based on their eigenvectors to provide a high-quality splitting of clean and corrupted examples (Kim et al. 2021). By leveraging the high-order topological information, most clean data are collected by a topological filter (Wu et al. 2020). (Wei et al. 2022) proposed a selection strategy, self-filtering (SFT), to filter noisy examples by utilizing their fluctuations. Nevertheless, these filters suffer from the overcleaning issue that overly removes even the true-labeled samples.

Although it has been verified that both correction and filtering are effective in dealing with label noises, how to choose the cleaning mode (correction or filtering) is still an unresolved critical problem. It involves their effectiveness, applicability, hyper-parameter sensitivity, etc.

Theoretical Results

In this section, we consider the random label noise problem in binary classification. The following analysis is based on the known noise probability and noise rate.

Notations

Suppose a classification dataset $D = \{(x_i, y_i)\}_{i=1}^N$ with label noise is drawn according to distribution $\mathcal{D}$. $\tilde{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ denotes the corresponding noise-free dataset, where $\tilde{y}_i$ is the ground-truth label. Let N, N_A, N_B be the sample sizes of datasets D, D_A, D_B from $\mathcal{D}$, respectively.

Definition 1 (Label noise)**.** *There exists label noise for the i-th sample if y_i is unequal to the ground-truth label $\tilde{y}_i$. Dataset D is noisy if there is at least one label noise in D.*

Definition 2 (Noise probability)**.** *The label noise probability*

$$p_i = \mathbf{Pr}(y_i \neq \tilde{y}_i). \tag{1}$$

Definition 3 (Noise rate)**.** *The label noise rate of dataset D*

$$\eta = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \neq \tilde{y}_i), \tag{2}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

Let η, η_A, η_B be the noise rates of D, D_A, D_B, respectively.

Definition 4 (Empirical error)**.** *The real empirical error of hypothesis $h(\cdot)$ on D is*

$$\hat{\mathcal{R}}_E(h; D) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(h(x_i) \neq y_i). \tag{3}$$

The true empirical error of hypothesis $h(\cdot)$ on D is

$$\mathcal{R}_E(h;D) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(h(x_i) \neq \tilde{y}_i). \tag{4}$$

It holds on a clean dataset that $\mathcal{R}_E(h,D) = \hat{\mathcal{R}}_E(h,D)$.

Noisy labels can be cleaned by correction and filtering at the data level. Noise correction refers to changing the wrong labels to the true labels. And noise filtering aims to exclude the samples with noisy labels.

Definition 5 (Cleaning size)**.** *The cleaning size (k) is the number of samples to be corrected or filtered. It refers to the number of corrected labels in noisy label correction and the number of removed samples in noise filtering.*

Definition 6 (Cleaned dataset)**.** *Let $\{(x_{(i)}, y_{(i)})\}_{i=1}^{N}$ be a dataset in which samples are ranked by a noise-related indicator, such as the noise probability (in descending order). The corrected dataset with cleaning size k is denoted by*

$$D^c = \{(x_{(i)}, y^c_{(i)})\}_{i=1}^{k} \cup \{(x_{(i)}, y_{(i)})\}_{i=k+1}^{N}. \tag{5}$$

The filtered dataset with cleaning size k is denoted by

$$D^f = \{(x_{(i)}, y_{(i)})\}_{i=k+1}^{N}. \tag{6}$$

The dataset with a single sample correction is

$$D^c_i = \{(x_{(j)}, y^c_{(j)})\}_{j=i} \cup \{(x_{(j)}, y_{(j)})\}_{j\neq i}. \tag{7}$$

The dataset with a single sample filtering is

$$D^f_i = D \backslash (x_{(i)}, y_{(i)}). \tag{8}$$

Their noise rates are denoted by $\eta^c, \eta^f, \eta^c_i, \eta^f_i$, respectively.

Error Bound in Learning with Label Noise

Definition 7 (Effectiveness)**.** *Let D_A, D_B be any two cleaned (corrected/filtered) sets of D. D_A is more* effective *than D_B, denoted by $D_A \succ D_B$ or $D_B \prec D_A$, if a given hypothesis/model trained on D_A has a lower error bound.*

Theorem 1 (Generalization error bound in learning with label noise)**.** *Let $\mathcal{H}$ be a family of functions with VC-dimension d. $\mathcal{D}$ denotes the distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over dataset D of size N with random label noise, the expected risk $\mathcal{R}(h;\mathcal{D})$ has the following upper bound for any $h \in \mathcal{H}$:*

$$\mathcal{R}(h;\mathcal{D}) \leq \hat{\mathcal{R}}_E(h;D) + [1 - 2\hat{\mathcal{R}}_E(h;D)]\cdot\eta + \epsilon(D,\mathcal{H},\delta), \tag{9}$$

where η denotes the noise rate of D, and

$$\epsilon(D,\mathcal{H},\delta) = \sqrt{[8d\log(2eN/d) + 8\log(4/\delta)]/N}. \tag{10}$$

The error bound increases with the noise rate and it becomes a conventional bound at $\eta = 0$. Hence it can be taken as a generalization of the classical bound. In addition, the error bound will be underestimated once the label noise is ignored. $\epsilon(D)$ is short for $\epsilon(D,\mathcal{H},\delta)$. By Theorem 1, the error bounds of label correction and filtering with cleaning size k are

$$\mathcal{B}^c_k = \hat{\mathcal{R}}_E(h;D) + [1 - 2\hat{\mathcal{R}}_E(h;D)]\cdot\eta^c + \epsilon(D^c), \tag{11}$$

$$\mathcal{B}^f_k = \hat{\mathcal{R}}_E(h;D) + [1 - 2\hat{\mathcal{R}}_E(h;D)]\cdot\eta^f + \epsilon(D^f), \tag{12}$$

where η^c, η^f denote the noise rates of corrected dataset D^c and filtered dataset D^f, respectively.

Theorem 2 (Effectiveness comparison)**.** *For any model,*

$$N_A > N_B, \eta_A < \eta_B \Rightarrow D_A \succ D_B, \tag{13}$$

$$N_A = N_B, \eta_A < \eta_B \Rightarrow D_A \succ D_B, \tag{14}$$

$$N_A > N_B, \eta_A = \eta_B \Rightarrow D_A \succ D_B, \tag{15}$$

where $N_A = |D_A|$, $N_B = |D_B|$. η_A, η_B are noise rates.

It means that a larger sample size and/or a lower noise rate yield a lower error bound. When the cleaning size is given, it can be concluded from the second case of Theorem 2 that samples with larger noise probabilities should be cleaned first both for correction and filtering. The reason is that this choice is more likely to reduce the noise rate and the error bound as much as possible for a fixed cleaning size.

Effectiveness of Correction and Filtering

Theorem 3 (Effectiveness of label correction)**.**

$$\eta^c < \eta \Leftrightarrow D^c \succ D. \tag{16}$$

It indicates that the noise rate is required to be lower than before for an effective label correction.

Theorem 4 (Effectiveness of filtering)**.**

$$\eta^f < \eta - \frac{\epsilon(D^f) - \epsilon(D)}{1 - 2\hat{\mathcal{R}}_E(h;D)} \Leftrightarrow D^f \succ D. \tag{17}$$

It implies that the noise rate of the filtered dataset should be significantly reduced to compensate for sample removal.

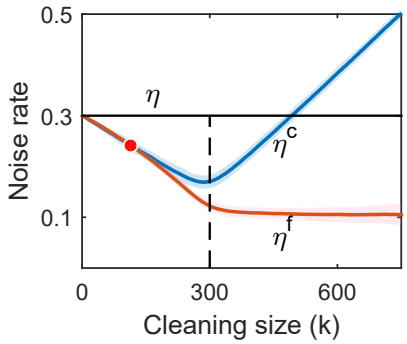


Figure 2: Noise rates of correction and filtering. 1000 noise probability values are directly generated from a beta mixture distribution, and they do not depend on any dataset or classifier in simulations. Pseudo-random numbers from two sub-distributions correspond to noisy and clean samples, respectively. The cleaning sequence is set partially by noise probability p_i descendingly. Initial noise rate $\eta = 0.3$, sample size $N = 1000$. The vertical dashed line represents the size of noises $\eta N = 300$. Results are averaged on 100 trails.

Figure 2 simulates the noise rates of correction and filtering. The noise rate of correction η^c decreases with the cleaning size on the left of the vertical dashed line, and it starts growing on the other side due to wrong corrections. The noise rate of filtering η^f decreases with the cleaning

size. It tends to be a small positive value owing to an imperfect noise probability estimate in the real situation.

Figure 3 shows the error bounds of correction and filtering. The bound of correction $\mathcal{B}_k^c$ is slightly lower than that of filtering $\mathcal{B}_k^f$ when the cleaning size is smaller than the size of noises (ηN). While the filtering is clearly more effective than the correction ($\mathcal{B}_k^f < \mathcal{B}_k^c$) when the cleaning size exceeds the size of noises, i.e. over-cleaning. The effective cleaning interval refers to the range of k such that the bound of correction/filtering is lower than the initial bound.The effective filtering interval is obviously wider than that of correction. It implies an imperfect cleaning size may have fewer negative effects on the bound of filtering than correction. In other words, filtering outperforms correction in **hyper-parameter insensitivity**. Besides, both optimal cleaning sizes ($\arg\min_k \mathcal{B}_k^c$ and $\arg\min_k \mathcal{B}_k^f$) are very close to the true size of noises. It means error bounds in (11), (12) could guide cleaners to avoid over-cleaning.

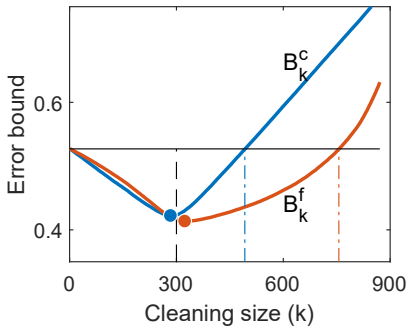


Figure 3: Error bounds of correction and filtering. These bounds are calculated by (11) and (12) based on pseudo-random numbers of noise probability. The sample size $N = 1000$, initial noise rate $\eta = 0.3$, and $\hat{\mathcal{R}}_E(h; D) = 0.1$. Function $\epsilon(\cdot)$ is calculated by (10), where VC-dimension $d = 10$, $\delta = 0.05$. Both bounds are averaged on 100 repetitions.

Comparison of Correction and Filtering

Correction and filtering are compared at the same cleaning size, i.e., the number of corrected labels is equal to that of removed samples.

Theorem 5 (Effectiveness comparison at the dataset level)**.**

$$\begin{cases} \eta^c < \eta^f + \frac{\epsilon(D^f)-\epsilon(D^c)}{1-2\hat{\mathcal{R}}_E(h;D)} \Leftrightarrow D^c \succ D^f, \\ \eta^f < \eta^c - \frac{\epsilon(D^f)-\epsilon(D^c)}{1-2\hat{\mathcal{R}}_E(h;D)} \Leftrightarrow D^c \prec D^f. \end{cases} \tag{18}$$

The choice of cleaning mode (correction/filtering) depends on the noise rate, cleaning size, and the (real) empirical error. From Theorem 5, we know $\eta^c = \eta^f \Rightarrow D^c \succ D^f$. It means the correction is superior to filtering in terms of **effectiveness** when the noise rate of the cleaned dataset is fixed. This is because the corrected dataset has more samples than the filtered dataset. It is consistent with the third case of Theorem 2. In Figure 2, it holds for the cleaning size $k = 108$ that $\eta^c = \eta^f$. And the bound of correction at $k = 108$ is lower than that of filtering in Figure 3.

Theorem 6 (Effectiveness comparison at the sample level)**.**

$$\begin{cases} p_i > p^T \Leftrightarrow D_i^c \succ D_i^f, \\ p_i < p^T \Leftrightarrow D_i^c \prec D_i^f, \end{cases} \tag{19}$$

where

$$p^T = \frac{N-\eta N-1}{N-2} - \frac{N(N-1)}{N-2} \cdot \frac{\epsilon(D_i^f) - \epsilon(D_i^c)}{1-2\hat{\mathcal{R}}_E(h;D)}. \tag{20}$$

Theorem 6 indicates the following **applicability** of cleaning modes. Compared to filtering, the correction is more effective in processing the label with a large noise probability ($p_i > p^T$) at the sample level. Filtering is suitable for handling samples with unconfident median noise probabilities ($0.5 < p_i < p^T$). Samples with low probabilities are confident and should be retained without any modification.

Suppose the samples have been ranked by the noise probability in descending order. We get the following criteria for **cleaning mode selection** from Theorems 5 and 6.

- In the early stage of noise cleaning, most noise probabilities of corrupted labels are large enough, and they could be easily corrected to reduce the noise rate and error bound without losing samples.
- In the later stage, it is so difficult to identify a true label noise that wrong corrections may raise the noise rate. As a conservative cleaning mode, filtering could produce a significantly lower noise rate than correction. Hence filtering is employed to avoid wrong corrections and alleviate the negative impact of over-cleaning.

The Proposed Fusion Cleaning Method

Noise Probability Estimation

The noise probability is estimated based on label confidence which can be produced by various classifiers (Wu et al. 2021). Assume the low label confidences are generated by a noise distribution and high confidences are mainly from another clean distribution. Then all label confidences can be fitted by a two-component ($T = 2$ in (21)) beta mixture model (BMM) but not a Gaussian mixture as the former better fits the skew toward zero confidence from noisy samples. The probability density function (PDF) of BMM

$$\rho(c) = \sum_{t=1}^{T} \lambda_t \rho(c|t), \tag{21}$$

where λ_t is the coefficient of the sub-distribution with PDF

$$\rho(c|t) = \frac{\Gamma(\alpha_t + \beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)} c^{\alpha_t - 1}(1-c)^{\beta_t - 1}, \tag{22}$$

$\Gamma(\cdot)$ is the Gamma function, and α_t, β_t are parameters of a beta distribution that can be determined by the expectation maximization (EM) method. Specifically, latent variable $\gamma_t(c)$ is defined to be the posterior probability of the confidence c being generated by the t-th sub-distribution. In the E-step, parameters $\lambda_t, \alpha_t, \beta_t$ are fixed, and the latent variable is updated by the Bayes formula

$$\gamma_t(c_i) = \frac{\lambda_t \rho(c_i|t)}{\sum_{j=1}^{2} \lambda_j \rho(c_i|j)}. \tag{23}$$

In the M-step, for fixed $\gamma_t(c_i)$, parameters α_t, β_t are estimated by the method of moments (1st and 2nd moments):

$$\hat{\alpha}_t = \frac{\bar{c}_t^2(1-\bar{c}_t)}{s_t^2 - \bar{c}_t^2} - \bar{c}_t, \hat{\beta}_t = \frac{\hat{\alpha}_t(1-\bar{c}_t)}{\bar{c}_t}, \tag{24}$$

where the weighted sample moments

$$\bar{c}_t = \frac{\sum_{i=1}^N \gamma_t(c_i)\cdot c_i}{\sum_{i=1}^N \gamma_t(c_i)}, s_t^2 = \frac{\sum_{i=1}^N \gamma_t(c_i)\cdot c_i^2}{\sum_{i=1}^N \gamma_t(c_i)}. \tag{25}$$

Coefficient λ_t is updated by

$$\lambda_t = \frac{1}{N}\sum_{i=1}^N \gamma_t(c_i). \tag{26}$$

The E-step and M-step are iterated until convergence or reach a maximum number of iterations. Finally, the label noise probability is estimated by

$$\hat{p}_i = \frac{\lambda_1 \rho(c_i|t=1)}{\sum_{t=1}^2 \lambda_t \rho(c_i|t)}, \tag{27}$$

where $\rho(c_i|t=1)$ denotes the PDF of noise sub-distribution with a smaller confidence mean.

The label confidences in both binary and multi-class classifications usually have a double-peak distribution. The main difference lies in the locations of peaks, e.g., the confidence of a clean label in binary classification is generally larger than that in a multi-class task. That means its distributions with different class numbers have no essential difference. Besides, the noise probability can be taken as a soft measurement of a binary determination (clean/noisy) problem, and it has been normalized by the EM estimate. Hence the above noise probability estimate is applicable to noise cleaning in multi-class classification.

Fusion of Label Correction and Filtering

Based on the comparison of correction and filtering, it is proper to correct high-probability noisy labels and remove unconfident samples. Then a fusion cleaning algorithm, called the Fusion of Correction and Filtering (FCF), is proposed in Algorithm 1. Labels with larger noise probabilities are corrected in the first round (steps 2-7) with cleaning size

$$k_1 = \arg\min_k \mathcal{B}_k^c = \arg\min_k \eta_k^c. \tag{28}$$

Samples with unconfident labels in the corrected set are removed in the second round (steps 8-13) with cleaning size

$$k_2 = \arg\min_k \mathcal{B}_k^f. \tag{29}$$

Note that the cleaning size is in a finite range of integers, the optimal solution can be found easily. The cleaned dataset and classifier are updated simultaneously. Then the confidence and noise probability would be generated more accurately as the training and cleaning go on. Besides, the noise probability estimation, label correction, and classifier may be implemented in various ways. Thus it can be taken as a general framework for data cleaning and robust modeling.

Algorithm 1: A Fusion cleaning of Correction and Filtering

Input: Noisy dataset $D = \{(x_i, y_i)\}_{i=1}^N$
Output: Cleaned dataset D^f, the final classifier

1: Train the classifier with D in the first Ep_0 epochs;
2: **for** epoch=Ep_0+1 **to** Ep_0+Ep_1 **do**
3: Generate/Updata the original label confidence c_i;
4: Compute the label noise probability via EM estimate;
5: Compute the error bound of correction $\mathcal{B}_k^c$ by (11), and find the optimal cleaning size k_1 by (28);
6: The k_1 labels with larger p_i in D are corrected and the classifier is trained on the cleaned dataset D^c.
7: **end for**
8: **for** epoch=Ep_0+Ep_1+1 **to** Ep_0+Ep_1+Ep_2 **do**
9: Generate/Updata the corrected label confidence c_i;
10: Compute the noise probability p_i via EM estimate;
11: Compute the error bound of filtering on the corrected dataset $\mathcal{B}_k^f$ by (12), and find k_2 by (29);
12: The k_2 samples with larger p_i in D^c are removed, and the classifier is trained on the cleaned dataset D^f.
13: **end for**

Multiclass Implement

Theorem 2 suggests that we should focus on samples with larger noise probabilities in label cleaning. In binary classification, a large-noise-probability label is changed to the other class in correction. However, the purified label may be uncertain in the multiclass task. As a result, the correction on the label with the largest noise probability may not produce the most reduction of noise rate. To address this issue, the key indicator in correction becomes the noise probability reduction in the multiclass task. It means we need to calculate the noise probability sets before and after correction, denoted by $\{p_i^0\}_{i=1}^N$ and $\{p_i^c\}_{i=1}^N$. Both sets are sorted descendingly by the reduction ($\Delta p_i^c = p_i^0 - p_i^c$), then we obtain the ordered sets $\{p_{(i)}^0\}_{i=1}^N$ and $\{p_{(i)}^c\}_{i=1}^N$. The noise rate of corrected dataset with cleaning size k

$$\hat{\eta}_k^c = \frac{1}{N}\left(\sum_{i=1}^k p_{(i)}^c + \sum_{i=k+1}^N p_{(i)}^0\right). \tag{30}$$

The noise rate of filtered dataset with cleaning size k

$$\hat{\eta}_k^f = \frac{1}{N-k}\sum_{i=k+1}^N p_{(i)}, \tag{31}$$

where $p_{(i)}$ is the ordered noise probability before filtering.

A good property of the above noise rate estimates is both optimal cleaning sizes in (28) and (29) are insensitive to the bias of noise probability estimation. Suppose that all probabilities have a constant bias. Then $p'_{(i)} = p_{(i)} + \varepsilon \Rightarrow \hat{\eta}_k^{f'} = \hat{\eta}_k^f + \varepsilon$. By (12), we know $\mathcal{B}_k^{f'} = \mathcal{B}_k^f + [1 - 2\hat{\mathcal{R}}_E(h;D)]\cdot\varepsilon$. Thus the optimal cleaning size $k_1' = k_1$.

Experiments

In this section, we empirically evaluate our proposed method on datasets with label noise.

Dataset	Method	Symmetric Noise				Asymmetric Noise		
		20%	40%	60%	80%	20%	30%	40%
CIFAR-10	Standard	85.7±0.5	81.8±0.6	73.7±1.1	42.0±2.8	88.0±0.3	86.4±0.4	84.9±0.7
	Bootstrap	86.4±0.6	82.5±0.1	75.2±0.8	42.1±3.3	88.8±0.5	87.5±0.5	85.1±0.3
	Joint-Opt	89.8±0.8	88.6±0.8	85.6±0.5	65.9±0.3	92.1±0.6	91.3±0.5	90.2±0.6
	Co-teaching	89.2±0.3	86.4±0.4	79.0±0.2	22.9±3.5	90.0±0.2	88.2±0.1	78.4±0.7
	MW-Net	90.1±0.7	86.4±0.5	81.6±0.2	64.8±0.6	92.0±0.7	91.3±0.5	90.9±0.4
	SELFIE	90.2±0.3	86.3±0.3	81.2±0.3	63.5±0.4	89.3±0.2	88.4±0.5	85.7±0.3
	AdaCorr	91.0±0.3	88.7±0.5	81.2±0.4	49.2±2.4	92.2±0.1	91.3±0.3	89.2±0.4
	TopoFilter	90.2±0.2	87.2±0.4	80.5±0.4	45.7±1.0	90.5±0.2	89.7±0.3	87.9±0.2
	MSLC	93.4±0.1	91.2±0.2	87.3±0.1	68.9±0.5	94.1±0.2	93.6±0.3	92.5±0.3
	FINE	90.8±0.2	87.6±0.2	81.0±0.4	69.4±1.2	92.2±0.1	91.6±0.2	89.5±0.2
	SFT	92.6±0.3	89.5±0.3	87.1±0.2	66.2±0.8	91.5±0.3	90.4±0.5	89.9±0.5
	CTO	88.1±0.3	85.1±0.2	80.3±0.1	67.2±2.4	85.3±0.2	82.3±0.5	80.3±0.8
	FCF(ours)	**93.7±0.3**	**91.8±0.4**	**87.6±0.3**	**76.3±2.8**	**94.4±0.3**	**94.1±0.4**	**93.6±0.6**
CIFAR-100	Standard	56.5±0.7	50.4±0.8	38.7±1.0	18.4±0.5	57.3±0.7	52.2±0.4	42.3±0.7
	Bootstrap	56.2±0.5	50.8±0.6	37.7±0.8	19.0±0.6	57.1±0.9	53.0±0.9	43.0±1.0
	Joint-Opt	60.1±0.9	56.8±0.7	47.7±0.4	17.4±0.5	66.7±0.3	63.4±0.6	59.3±0.6
	Co-teaching	64.8±0.2	60.3±0.4	46.8±0.7	13.3±2.8	63.6±0.4	58.3±1.1	48.9±0.8
	MW-Net	68.4±0.7	64.8±0.3	55.0±0.3	19.2±0.2	66.7±0.4	63.2±0.5	59.5±0.4
	SELFIE	67.2±0.3	61.3±0.4	53.1±0.6	20.4±0.4	65.2±0.2	62.8±0.4	58.7±0.5
	AdaCorr	67.8±0.1	60.2±0.8	46.5±1.2	24.6±1.1	68.3±0.2	61.1±0.5	49.8±0.7
	TopoFilter	65.6±0.3	62.0±0.6	47.7±0.5	20.7±1.2	68.0±0.3	66.7±0.6	62.4±0.2
	MSLC	72.0±0.3	68.7±0.2	60.3±0.3	20.5±1.8	70.2±1.2	69.7±0.8	69.2±0.6
	FINE	69.1±0.2	64.7±0.5	62.3±0.4	25.6±1.3	68.5±0.3	64.9±0.7	61.7±1.0
	SFT	72.0±0.3	69.7±0.3	60.4±0.4	25.2±1.7	71.2±0.3	70.1±0.3	69.3±0.4
	CTO	68.2±0.3	61.7±0.4	51.8±0.2	20.6±2.5	62.8±0.3	56.4±0.5	44.6±1.2
	FCF(ours)	**74.3±0.4**	**69.8±0.7**	**63.0±0.6**	**26.7±2.3**	**76.3±0.3**	**74.8±0.8**	**73.4±1.5**

Table 1: Test accuracy (%) comparison. Best results are marked in bold. All results are averaged on three independent trials.

Datasets and label noises. We implement experiments on CIFAR-10 and CIFAR-100 under different types and levels of noise. Both datasets consist of 50k training images and 10k test images of size 32×32. Following the previous setups (Kim et al. 2021), we manually generate two types of noisy labels: symmetric and asymmetric. Symmetric: a given proportion of labels are changed to one of the other class labels uniformly. Asymmetric: a label is corrupted only to a specific similar class, e.g., truck→automobile, bird→airplane, deer→horse, cat↔dog for CIFAR-10. For CIFAR-100, the label flip only happens in each super-class.

Baselines. Our method is compared to diverse baselines. *Standard* trains the deep network with cross-entropy. *Bootstrap* (Reed et al. 2014) and *MW-Net* (Shu et al. 2019) aim to generate robust classifiers by modified loss function and sample weighting. *Joint-Opt* (Tanaka et al. 2018), *AdaCorr* (Zheng et al. 2020), *MSLC* (Wu et al. 2021), *CTO* (Li and Sun 2022) are representative label correctors. *Co-teaching* (Han et al. 2018), *SELFIE* (Song, Kim, and Lee 2019), *TopoFilter* (Wu et al. 2020), *FINE* (Kim et al. 2021), *SFT* (Wei et al. 2022) belong to sample selection/filtering methods. For a fair comparison, we compare them without mixup augmentation and semi-supervised learning methods.

Experimental setup. We use ResNet-34 as the backbone, and train the network for 180 epochs (80 in warm-up, 40 in correction, 60 in filtering) on CIFAR-10/100 for our proposed FCF. We use an SGD optimizer with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 100. The initial learning rate is set as 0.1 with decaying by a factor of 0.1 in epochs 80, 100, and 150, respectively. The experimental results are listed in Table 1. It can be observed that FCF outperforms other methods across all the label noise settings on CIFAR-10 and CIFAR-100.

Label purity of cleaned dataset. From Figure 4, both test accuracy and label purity have positive jumps at the beginning of correction and filtering stages. This verifies their effectiveness in improving label quality and generalization performance. Besides, the correction size (k_1) is close to the true noise level.

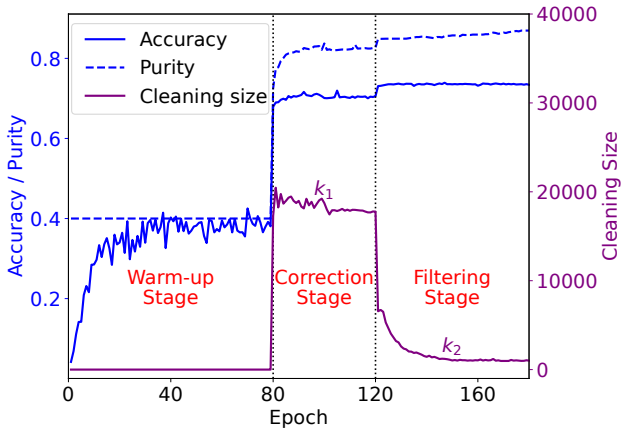


Figure 4: Accuracy and label purity of FCF on CIFAR-100 with 40% symmetric noise. Label purity denotes the correctness of all labels in the cleaned dataset.

Ablation Study. Table 2 lists the average accuracies with different cleaning modes. Both correction and filtering could improve the test accuracy and the former is more effective. Our fusion cleaning mode (FCF) has the best performance in different noise environments.

Cleaning mode	Symmetric noise		Asymmetric noise	
	20%	40%	20%	40%
Without cleaning	47.6	45.3	51.9	37.7
Correction	72.1	68.7	72.7	69.3
Filtering	69.4	65.1	70.1	65.5
Corr. & Filt.(FCF)	**74.3**	**69.8**	**76.3**	**73.4**

Table 2: Accuracy on CIFAR-100.

Figure 5 shows the increment of confusion matrix in FCF noise cleaning. All diagonal values are positive and the others are nonpositive. It means correction and filtering in FCF improve the label purity twice. The filtering effect is relatively more significant in dealing with confusable classes, such as dog and cat (5↔3), bird and airplane (2→0).

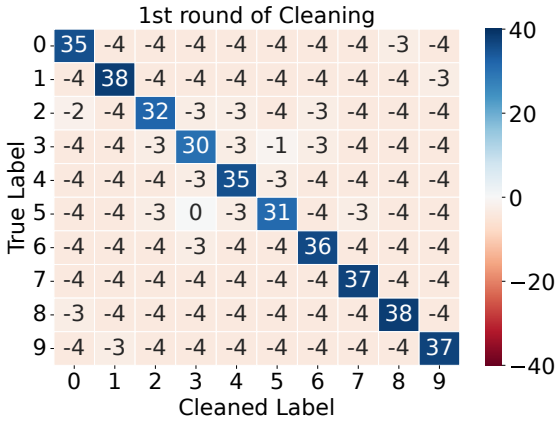
Data cleaning on real CIFAR-10 test set. Our proposed algorithm is validated on original CIFAR-10 test set (10k). Some images with label issues are displayed in Figure 6. These low-confidence labels of top images are changed to the correct ones. While it is difficult to categorize the bottom images with median confidence due to the low quality. It is better to remove samples than to change the label to an unconvincing one. Besides, both average cleaning sizes (k_1=15, k_2=214) on the test set are far less than those under artificial noise as displayed in Figure 4. It verifies the adaptability of our theories and method to some extent.

Conclusion

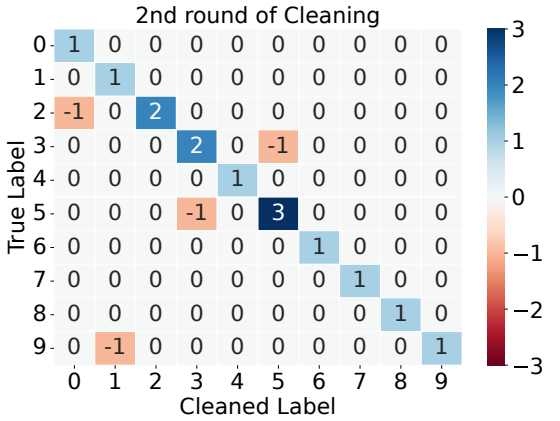
In this paper, we rebuild the error bound under label noise which can be viewed as a generalized form of the conventional bound. The correction and filtering techniques are compared in terms of effectiveness, applicability, and hyper-parameter insensitivity from an error-bound perspective. Then we summarize the criteria for selecting the cleaning modes. Unlike existing methods with a single cleaning mode, the proposed FCF combines the advantages of correction and filtering to deal with different types of suspicious labels. And FCF is a general cleaning framework that can be integrated with other noise probability estimates and label correctors. Experimental results show that FCF significantly improves the label quality and achieves state-of-the-art performance. As an error-bound-guided cleaning method, FCF expands the applicability of statistical learning theory in designing practical and effective algorithms.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Nos. 62276161, U21A20513, 12226004, 62076154, 61906113, 61721002) and the Key R&D program of Shanxi Province (No. 202202020101003).



(a) Correction



(b) Filtering

Figure 5: Confusion matrix increment $M_A - M_B$ (%) on CIFAR-10 under 40% symmetric noise, where M_B, M_A are confusion matrices before and after cleaning, respectively.

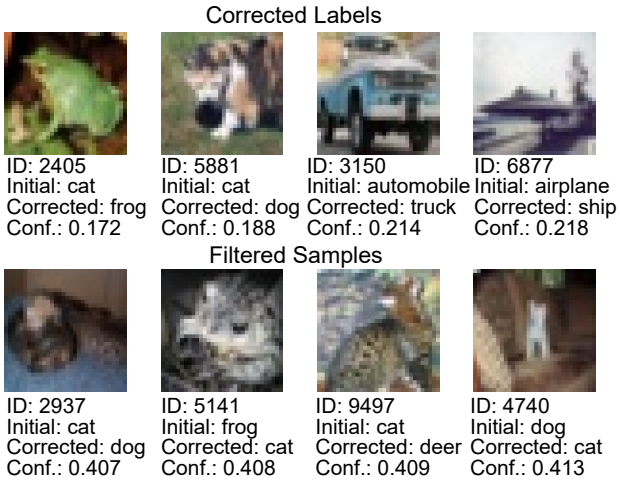


Figure 6: Cleaned examples in CIFAR-10 test set. Top image labels are corrected and bottom images are filtered by FCF. Labels before and after correction are listed below images.

References

Frenay, B.; and Verleysen, M. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5): 845–869.

Garcia, L. P. F.; Lorena, A. C.; Matwin, S.; and de Carvalho, A. 2016. Ensembles of label noise filters: A ranking approach. *Data Mining and Knowledge Discovery*, 30(5): 1192–1216.

Han, B.; Tsang, I.; Chen, L.; Zhou, J.; and Yu, C. 2019. Beyond majority voting: A coarse-to-fine label filtration for heavily noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 30(12): 3774–3787.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 8527–8537.

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-adaptive training: Beyond empirical risk minimization. In *Advances in Neural Information Processing Systems*, volume 33, 19365–19376.

Jiang, G.; Wang, W.; Qian, Y.; and Liang, J. 2021. A unified sample selection framework for output noise filtering: An error-bound perspective. *Journal of Machine Learning Research*, 22(18): 1–66.

Kim, T.; Ko, J.; Cho, S.; Choi, J.; and Yun, S. 2021. FINE samples for learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 34, 24137–24149.

Li, J.; Socher, R.; and Hoi, S. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*.

Li, J.; and Sun, H. 2022. Correct twice at once: Learning to correct noisy labels for robust deep learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5142–5151.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. DISC: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24070–24079.

Northcutt, C.; Athalye, A.; and Mueller, J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 34.

Northcutt, C.; Jiang, L.; and Chuang, I. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70: 1373–1411.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Sáez, J. A.; Galar, M.; Luengo, J.; and Herrera, F. 2016. INFFC: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control. *Information Fusion*, 27: 19–32.

Sharma, K.; Donmez, P.; Luo, E.; Liu, Y.; and Yalniz, I. Z. 2020. Noiserank: Unsupervised label noise reduction with dependence models. In *European Conference on Computer Vision*, 737–753.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 1919–1930.

Shu, J.; Yuan, X.; and Meng, D. 2023. CMW-net: An adaptive robust algorithm for sample selection and label correction. *National Science Review*, 10(6): nwad084.

Song, H.; Dai, R.; Raskutti, G.; and Barber, R. F. 2020. Convex and non-convex approaches for statistical inference with class-conditional noisy labels. *Journal of Machine Learning Research*, 21(168): 1–58.

Song, H.; Kim, M.; and Lee, J.-G. 2019. SELFIE: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, 5907–5915.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2022.3152527.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5552–5560.

Wei, Q.; Sun, H.; Lu, X.; and Yin, Y. 2022. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*, 516–532.

Wu, P.; Zheng, S.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. A topological filter for learning with label noise. In *Advances in Neural Information Processing Systems*, 21382–21393.

Wu, Y.; Shu, J.; Xie, Q.; Zhao, Q.; and Meng, D. 2021. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10388–10396.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, 7597–7610.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zheng, S.; Wu, P.; Goswami, A.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, 11447–11457.