Pattern Recognition Letters 184 (2024) 239–245

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Decoding class dynamics in learning with noisy labels

Albert Tatjer [a,1], Bhalaji Nagarajan [a,*,1], Ricardo Marques [a,b,2], Petia Radeva [a,2]

[a] Department de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007, Barcelona, Spain
[b] Computer Vision Center (CVC), Campus UAB, Edifici O, Cerdanyola del Vallès, 08193, Barcelona, Spain

ARTICLE INFO

Editor: Maria De Marsico

MSC:
41A05
41A10
65D05
65D17

Keywords:
Learning with Noisy Labels
Label Noise Modelling
Class Dynamics

ABSTRACT

The creation of large-scale datasets annotated by humans inevitably introduces noisy labels, leading to reduced generalization in deep-learning models. Sample selection-based learning with noisy labels is a recent approach that exhibits promising upbeat performance improvements. The selection of clean samples amongst the noisy samples is an important criterion in the learning process of these models. In this work, we delve deeper into the clean-noise split decision and highlight the aspect that effective demarcation of samples would lead to better performance. We identify the Global Noise Conundrum in the existing models, where the distribution of samples is treated globally. We propose a per-class-based local distribution of samples and demonstrate the effectiveness of this approach in having a better clean-noise split. We validate our proposal on several benchmarks — both real and synthetic, and show substantial improvements over different state-of-the-art algorithms. We further propose a new metric, classiness to extend our analysis and highlight the effectiveness of the proposed method. Source code and instructions to reproduce this paper are available at https://github.com/aldakata/CCLM/

1. Introduction

Recent breakthroughs in Deep Learning (DL) have led to remarkable achievements across various intricate tasks and have surpassed human-level performances. The high performance of these models hinges on high-quality large-scale datasets, which, however, involves substantial cost and labour-intensive efforts [1]. Despite the substantial efforts, mislabelling remains unavoidable due to the data complexity and potential human annotation errors [2]. Weakly supervised approaches such as crowdsourcing [3] and web-crawling [4] are particularly susceptible to label noise due to human participation. Introduction of noise in the assigned labels thus becomes inevitable [5,6]. Training Deep Neural Networks (DNNs) with noisy labels leads to a significant decline in their generalization performance [7]. Over-parameterization [8] and DNN's strong memorization capability [9] often lead to over-fitting to this label noise. Traditional regularization techniques, such as dropouts or weight decay, fall short of effectively mitigating this problem [9].

Learning with Noisy Labels (LNL) has been a long-studied problem ever since its inception in the last 1980s [10]. It holds a crucial position within the DL community due to its critical role in developing robust models capable of handling label noise. Several strategies have been proposed to create noise-robust DNNs [11]. Making loss function modifications [12,13], creating new robust loss functions [14,15], adding regularization [16,17], reweighting samples [18,19], label correction [20,21], label aggregation [22], learning jigsaw puzzles [23], and using topological structure information [24] have contributed to alleviating the impact of label noise.

Sample selection-based methods are the most straightforward and widely adopted LNL method. They work on identifying and selecting clean samples from the training samples and imposing different schemes in order to reduce the interference of noisy samples. Different criteria are used in sample selection-based methods, of which, the small-loss criterion is a popular scheme [25,26]. In this criterion, the small loss samples are treated as the clean samples [27]. The success of sample selection methods depends on the quality of the clean-noise separations and the criteria designed to configure the refined set [28].

Recent LNL methods have enhanced robustness by combining diverse semi-supervised learning strategies. **DivideMix** (DM) [26] is a prominent LNL benchmark algorithm based on the modelling of sample loss distribution. DM trained two 'peer' DNNs and modelled the loss distribution of each model using a Gaussian Mixture Model (GMM) in order to dynamically divide the clean and noisy samples. DM uses the

* Corresponding author.
E-mail addresses: albert.catalan-tatjer@student.uni-tuebingen.edu (A. Tatjer), bhalaji.nagarajan@ub.edu (B. Nagarajan), ricardo.marques@ub.edu (R. Marques), petia.ivanova@ub.edu (P. Radeva).
[1] Equal contribution.
[2] Equal supervision.

https://doi.org/10.1016/j.patrec.2024.04.012
Received 31 October 2023; Received in revised form 25 January 2024; Accepted 15 April 2024
Available online 19 April 2024
0167-8655/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

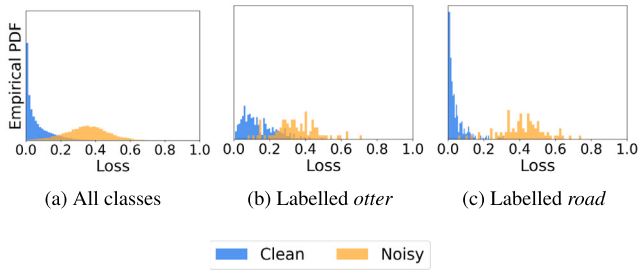

Fig. 1. Empirical loss distribution of the clean samples and the noisy samples after the warm-up phase (CIFAR-100 20% Sym. noise).

clean samples as a labelled set and noisy samples as an unlabelled set and employs MixMatch [29] to learn the samples in a semi-supervised manner. **Contrast to Divide** (C2D) [30], improved DM by applying self-supervised pre-training. One of the important design questions revolves around *the effective selection of clean and noisy samples* playing an important role as it directly affects the subsequent tasks in the training pipeline.

DNNs tend to learn the underlying patterns of the data first before gradually memorizing the samples [7]. Small-loss-based methods assume that the smaller loss samples have a lower likelihood of being affected by label noise. However, not all samples are learned in this fashion [31], which hinders noise detection in these methods, leading to *Global loss conundrum*. In our previous work published in IbPRIA 2023, **Class-Conditional Local noise Model** (CCLM) [32], we studied this global noise modelling and proposed a per-class label noise model. In this work, we extend our work on CCLM in several dimensions: (1) We provide an extended comprehensive review of the latest SoTA works that are related to our proposal (Section 2). (2) We extend CCLM to other SoTA LNL algorithms, namely, Contrast to Divide (Section 4). (3) We further expand the evaluations to two real noise benchmarks [6] along with the two synthetic noise benchmarks used in CCLM. (4) We provide a deeper analysis by including new metrics and extend the analysis of CCLM applied to both DM and C2D (Section 5.2). The obtained results highlight the effectiveness of CCLM in label noise modelling.

2. Related work

LNL is of significant research interest and several methods have been proposed to tackle the problem of label noise [11]. Using noise transition matrix [12,13], employing regularization methods such adding dropouts [17], gradient clipping [16], progressive early stopping [33], SANM [34] have alleviated the effect of label noise. Robust loss functions such as Symmetric cross-entropy learning [35] and Active Passive Loss [36] have increased the performance of models in the presence of label noise. Recently, contrastive learning methods have been used in the context of LNL algorithms. ChiMera [37], UNICON [38] and TCL [39] have been successful in LNL training. Additional knowledge, such as using uncertainty, help complement the learning process [40,41].

Sample selection methods work on selecting clean samples amongst the noisy samples and devise different strategies to reduce the impact of the label noise. They work on the intuition that less noisy data leads to more robust DNNs [7]. Different strategies exist in demarcating clean and noisy samples of which small-loss selection [25,27] is the most widely adopted. In this selection criterion, the small loss samples are treated as clean samples [27]. A prevalent issue of sample selection methods is the accumulation of errors by incorrect selection of samples. 'Peer' networks or multiple DNNs, such as Co-teaching [42], Co-teaching+ [43] and JoCoR [44] were used to reduce this confirmation bias.

Different **loss modelling strategies** exist in the literature, such as using a beta-distribution model [25] and using GMMs [25,26] to split the training samples into clean and noisy samples. UNICON [38] used a Jensen–Shannon divergence based criterion to maintain class-balancing in selecting the clean samples. SSS-Net [45] combined shadowed-sets theory with clustering based on loss-similarity to select the clean samples. Recent sample selection methods incorporated advanced strategies based on the confidence of the models. DISC [46] incorporated confidence of samples to devise a dynamic thresholding strategy. Gradient Switching Strategy [47] used gradient direction as a weighting mechanism for the high-confidence and uncertain samples. CoDis [48] used high discrepancy data to maintain the divergence of two networks so as to enable the models to mine hard clean samples. Label Confidence Incorporation framework [49] incorporated label confidence as a measure of label noise and in turn prioritized the training samples.

One of the most widely recognized LNL multi-network benchmark algorithms in the context of sample selection methods is **DivideMix** [26]. DM trains two 'peer' networks to train each other by modelling the per-sample loss distribution using GMMs. It uses an improved MixMatch [29] algorithm to train the models using the clean and noisy samples as labelled and unlabelled sets. Several methods have been since developed highlighting the potential pitfalls in its training pipeline and improving the potential drawbacks. Augment Descent [50] used two different augmentation schemes instead of a global data augmentation pipeline. Contrast to Divide [30] proposed a self-supervised pre-training to create better feature extractors during the warm-up phase. ProMix [51] improved the small-loss sample selection by also considering the confidence of the samples. SplitNet [52] used an additional network to predict the samples being clean or noisy. PNP [53] used different optimizations based on the type of the samples. Manifold DivideMix [54] employed the k-nearest algorithm to remove out-of-distribution samples and used an iterative method to find the clean and noisy samples. Bayesian DivideMix++ [55] added several components to the training pipeline to improve the robustness against DNN memorization and utilized uncertainty metrics to further enhance its effectiveness. ULC [56] used different uncertainty measurements in label correction for imbalanced LNL problems.

One of the patterns observed during the training of DNNs is that not all samples are learned in the same way. **Class-based conditional** strategies have been used to better model the training data. CMW-Net [57] used a metamodel to learn a weighting scheme based on the training class. A dynamic class-conditional weighting was used to improve the balance of losses in DivideMix [31]. CPC [58] considered the loss distribution in a heterogeneous fashion and used class-aware modulation to partition the clean and noisy data. LCCN [59] used a Bayesian framework to parameterize the noise transition. SNSCL [60] employed a stochastic noise-tolerated supervised contrastive learning framework to learn representations in fine-grained classes. On this front, CCLM [32] proposed a class-conditional label noise modelling instead of the global noise modelling used in DivideMix. CCLM was able to show how the local treatment of samples fared better with respect to the global treatment of loss distribution. Following this, we extend our findings of CCLM in this manuscript.

3. Rationale

The typical approach taken by the most performing algorithms in LNL is to actively detect the potentially noisy samples and treat these samples differently from the samples which are deemed clean [26]. A common procedure is to apply semi-supervised learning techniques [1] to make the most out of the information provided by samples (either clean or noisy). Given the differentiated treatment applied to clean and noisy samples, an accurate label noise detection is thus essential for a successful learning procedure in the presence of noisy labels. On the one hand, treating noisy samples as clean samples often leads the model to learn (and thus, to memorize) wrong information, which

can significantly hinder the final accuracy of the model. On the other hand, treating clean samples as noisy can lead the model to discard relevant (and correct) information, eventually introducing mistakes in the learning process which degrade the learning outcome. To deal with this problem, LNL-based algorithms usually rely on a *noise model*, the goal of which is to capture a *global* statistical characterization of the samples' label noise. During training, this model is used to assign a probability of being clean (or noisy) to each sample in the training set, based on which the separation between clean and noisy samples is made. Typically, the noise model relies on a noisy proxy metric (i.e. the loss), leveraging the observation that noisy samples are harder to learn and thus tend to have larger loss values [25]. We provide an in-depth analysis of the *global noise model* approach and show that such an approach is locally suboptimal.

3.1. Global noise modelling

When tackling the problem of label noise detection, many existing approaches rely on the principle that noisy samples are more difficult to learn [61]. Therefore, when a neural network is used to evaluate the loss associated with each sample in the training set, noisy samples are statistically expected to yield larger losses, thus enabling the separation between clean and noisy samples. To formalize this process, let us consider a dataset $\mathcal{D}$ of n labelled samples such that $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, with x_i being the ith sample image and y_i its corresponding one-hot encoded label. Let us also consider a model m_Θ, where Θ are the model parameters, and a loss l function such that the set $\mathcal{L}$ of individual losses l_i for each sample $x_i \in \mathcal{D}$ is given by $\mathcal{L} = \{l_i = l\left(m_\Theta(x_i), y_i\right)\}_{i=1}^{n}$. An example of a histogram of $\mathcal{L}$ considering the cross-entropy loss is shown in Fig. 1(a). We can appreciate a clearly bimodal distribution, where the noisy samples, shown in orange, exhibit a characteristic loss generally larger than that of the clean samples, shown in cyan. Once the set $\mathcal{L}$ of all losses is computed, it is fed to the global noise model to produce a probability of a given sample being correctly (or wrongly) labelled. To this end, Li et al. [26] fit a 2-component Gaussian Mixture Model (GMM) to $\mathcal{L}$. Then, they retain the GMM component g with smaller mode, and use it to compute the clean probability of the ith given by $\omega_i := p(g|l_i)$. Lastly, the dataset $\mathcal{D}$ is partitioned into a clean $\mathcal{D}_C$ and a noisy set $\mathcal{D}_N$ given by:

$$\begin{aligned} \mathcal{D}_C &= \{(x_i, y_i)|\omega_i \geq \tau\}_{(x_i,y_i)\in\mathcal{D}}, \\ \mathcal{D}_N &= \{(x_i, y_i)|\omega_i < \tau\}_{(x_i,y_i)\in\mathcal{D}}, \end{aligned} \tag{1}$$

with τ being a threshold over the clean probability ω_i, and $l_i = l(m_\theta(x_i), y_i)$ as previously stated.

Drawbacks of global noise modelling. Modelling noise as a global process relies on the hypothesis that the per-sample loss behaves consistently across classes. Indeed, in this approach, the samples' losses are assumed to be i.i.d. realizations of a bi-modal loss distribution with 2 populations, where the population with a smaller mean corresponds to the clean samples, and the other corresponds to the noisy samples. Nevertheless, since different classes can have different characteristic losses, noisy samples of a given class might have a characteristic loss distributed differently from noisy samples of another class. This can potentially cause noisy samples to have a loss that falls within the limits of the low mean population, or, conversely, clean samples have a loss value around the mean of the high-loss values population. This observation is confirmed by comparing the per-sample loss distribution for the 'otter' class (Fig. 1(b)) with that of the 'road' class (Fig. 1(c)). Indeed, the loss distribution of both classes differs significantly from the global loss distribution (Fig. 1(a)). Consequently, the global noise model, although globally optimal, is clearly locally suboptimal regarding each individual class. We identify this situation as an obstacle to an accurate division between clean and noisy samples, which, in turn, can have significant negative effects on the training outcome. We hypothesize that better clean/noisy data partitions and increased model accuracy can be achieved using a class-conditional approach to the problem of noise label detection, therefore relaxing the aforementioned strong i.i.d. assumption on the samples' loss distribution. To test our hypothesis, we propose a class-conditional noise label model, aimed at overcoming the aforementioned limitations of global label noise modelling.

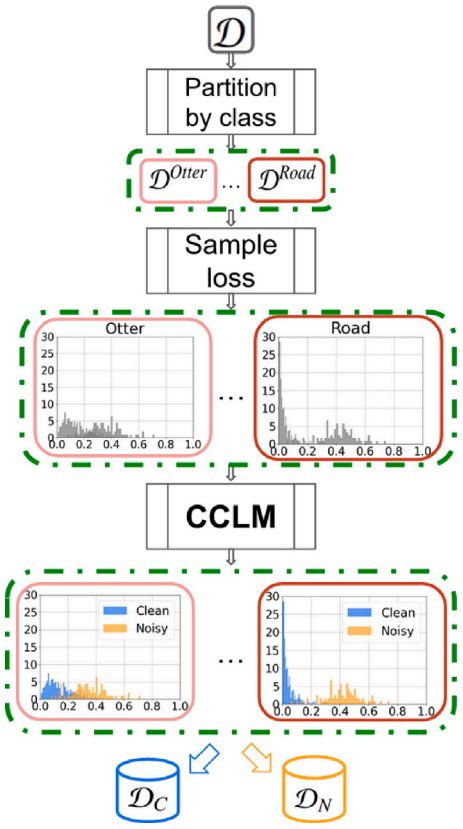

Fig. 2. CCLM partitions the dataset by the class labels and then fits a GMM on the sample loss of each subset. Subsequently, the resulting distribution combined with an adjustable threshold determines the Clean-Noisy partition.

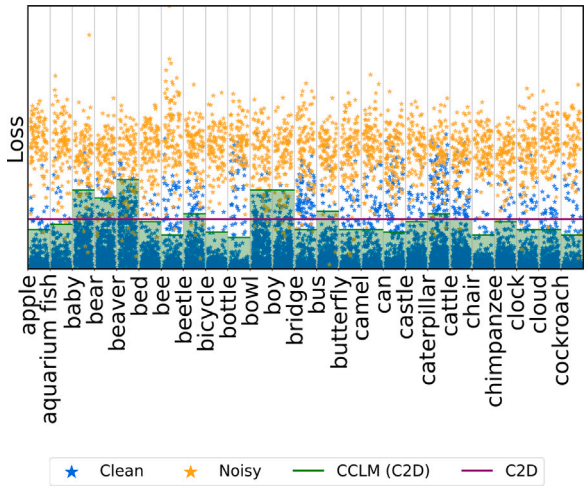

Fig. 3. Empirical sample loss distribution for 25 random classes of CIFAR-100 with Sym. 20% noise. The loss was computed at epoch 30 of the training procedure, using the C2D algorithm.

4. Proposed methodology

In this section, we propose a new label noise model which, as opposed to the typical approaches in LNL, is able to locally adapt to the particularities of the noise features of each class present in the dataset. This is achieved by explicitly including the information brought by the label class in the noise modelling process. As shown by our results, our method, which we label *Class-Conditional Local Noise Model* (CCLM), leads to a generalized improvement of the label noise detection task. Furthermore, CCLM can be coupled with widely used LNL algorithms such as DivideMix [26] or C2D [30]. Fig. 2 provides a global view of our proposed approach. The input noisy training set D is defined as $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, with x_i being the ith sample image and y_i its corresponding one-hot encoded label. The dataset D is then partitioned into J disjoint subsets, where J is the total number of classes, such that $D = \cup_{j=1}^{J} D^j$. Once the data is divided into classes, a 2-component GMM is fit to the per-sample loss of each subset D^j and, the GMM component with a lower mode is used to model the clean probability of the samples. Note that these results in a different distribution for each class, tailored to its specificity. Within each class, we then split between clean and noisy samples by thresholding the clean probability such that:

$$\begin{aligned} D_C &= \cup_{j=1}^{J} D_C^j = \cup_{j=1}^{J} \{(x_i, y_i) \mid \omega_i^j \geq \tau\}_{(x_i,y_i)\in D^j} \\ D_N &= \cup_{j=1}^{J} D_N^j = \cup_{j=1}^{J} \{(x_i, y_i) \mid \omega_i^j < \tau\}_{(x_i,y_i)\in D^j}, \end{aligned} \tag{2}$$

where τ is the same threshold used for all classes.

A comparison between our approach and the global noise modelling strategy is shown in Fig. 3. The horizontal axis represents a subset of 25 classes randomly chosen, while the vertical axis measures the per-sample loss. Clean samples are depicted as cyan dots, whereas noisy samples are depicted in orange. The split between clean and noisy samples for the global noise model is given by the magenta horizontal line (baseline): samples above the line are deemed noisy, whereas samples below the line are considered clean. This results in a division with a single decision boundary for all classes, effectively ignoring each class' particular noise features. The split between clean and noisy samples for our model is given by the green vs white areas corresponding to each class: samples within the green area are deemed clean, whereas samples within the white area are considered noisy. Note that, as opposed to the global noise model, our model provides a significantly different loss threshold for each class, even if the same probability threshold τ is used for all classes. This is because each class has its own clean probability function, so applying the same probability threshold for all classes results in a different loss threshold in practice. As shown next, our locally adapted split yields improvements both in the noise detection accuracy and in the final model classification accuracy when trained in the presence of noisy labels.

5. Experiments

Datasets. We evaluate the performance of CCLM on two synthetic noise benchmarks – CIFAR-10 [62] and CIFAR-100 [62] at different noise rates, as well as on two recent real noise benchmarks – CIFAR-10N [6] and CIFAR-100N [6]. All four benchmarks consist of 50,000 training samples and 10,000 test samples, where each image is of size 32×32. Following previous works [26,30], for the synthetic label noise, we validate our method on two types of noise — Symmetric (Sym) and Asymmetric (Asym). Symmetric noise is introduced by substituting the labels of a percentage of samples with random labels (all possible classes) drawn from a uniform distribution. Asymmetric noise is introduced by replacing the labels with "similar" classes (For example, 'Bird' and 'Airplane' in CIFAR-10). We vary the amount of injected noise (Sym. 20%, 50%, 80%, 90%, and Asym. 40%), following Patrini et al. [63] that is used as a standard experimental setting [25,26,30]. CIFAR-10N and CIFAR-100N use the training data of the corresponding CIFAR datasets with human-annotated real-world noisy labels. CIFAR-10N has five noisy label sets,[3] whereas CIFAR-100N has fine and coarse labels. We use the Noisy-Fine setting for our experiments. We evaluate all benchmarks using an 18-layer PreAct ResNet [64] as the backbone architecture. The experiment settings are presented in the Supplement.

Comparison methods. We compare the proposed CCLM against small-loss, sample-selection multi-network benchmark algorithms, DivideMix [26] and C2D [30]. Both methods follow a global noise model approach. CCLM (DM) corresponds to CCLM on DivideMix, whereas CCLM (C2D) corresponds to CCLM on C2D. We compare CCLM against DivideMix and C2D, as they are baseline frameworks for several algorithms [50,51,53,55].

5.1. Results

Following the experiments of DivideMix [26] and C2D [30], we report the best test accuracy (%) over all epochs (Best) and the average of the last 10 epochs (Last). The test set is used to provide insights into the generalization capabilities of the method. Long-time training may lead to performance degradation [65] and therefore, "last" is used as a measure of robustness [66,67]. Larger gaps between "best" and "last" correspond to overfitting.

Performance on synthetic noise benchmarks. Table 1 shows the test accuracy of CIFAR-10 and CIFAR-100 with different noise ratios. On both CIFAR datasets, CCLM (DM) outperforms DivideMix in all symmetric noise ratios. In the case of C2D, our method CCLM (C2D) works better on low-noise settings. However, the difference is marginal in high-noise settings. One of the possible reasons could be that the self-supervised pre-training induces a lot of class biases. In the high noise levels, it is hard to overcome such a strong trend with very few clean samples.

Performance on real noise benchmarks. We evaluate the performance of CCLM using two real-noise benchmarks, CIFAR-10N and CIFAR100-N (Table 2). In most noise levels of CIFAR-10N, CCLM fares better than DivideMix, achieving a significant performance gain of 1.3% in CIFAR-100N. However, the difference between CCLM (C2D) and C2D is marginal. Similar to the case of synthetic noise benchmarks, the small impact of CCLM on C2D can be explained by the effect of the pre-training phase used in C2D.

5.2. Analysis

We further analyse the behaviour of CCLM using two different metrics — Classiness and Noise Division Accuracy.

Classiness. Historically, statistical learning has been concerned with evaluation metrics and their potential for misinformation. For instance, model accuracy does not fare well in large class imbalances. We argue that LNL is one such vulnerable setting. As we can see in Table 1, on CIFAR-100 with 20% symmetric noise, DivideMix achieves a model accuracy of 77.3%. However, it is interesting to also understand the behaviour of LNL algorithms on different individual classes, as, not all classes are uniformly represented in a dataset (varying sample counts, sample difficulties, etc.).

In order to track this phenomenon, we introduce a new metric, called *classiness*. Classiness is defined as the standard deviation of the model accuracy over the classes, with a lower bound of zero and an unbounded upper bound. Let m and D be the model and dataset under evaluation, respectively. The accuracy μ of model m on the dataset D is given by $\mu = \mathrm{Acc}(m, D)$. Then the classiness of m on D can be defined as:

$$\mathrm{Classiness}(m, D) = \sqrt{\frac{1}{C}\sum_{j=1}^{C} \frac{N^j}{N} \left(\mathrm{Acc}(m, D^j) - \mu\right)^2}, \tag{3}$$

[3] Construction of label sets are detailed in CIFAR-N website.

Table 1
Synthetic noise performance comparison (test accuracy (%) on CIFAR datasets under different noise types). Bold indicates best performance.

Method		CIFAR-10					CIFAR-100				
		Symmetric				Asym.	Symmetric				Asym.
		20%	50%	80%	90%	40%	20%	50%	80%	90%	40%
DivideMix [26]	Best	96.1	94.6	93.2	76.0	**93.4**	77.3	74.6	60.2	31.5	72.2
	Last	95.7	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0	72.4
CCLM (DM)	Best	**96.5**	**95.6**	**93.7**	**83.6**	92.6	**79.5**	**76.4**	**61.1**	**33.5**	**75.4**
	Last	96.3	95.3	93.6	82.4	91.5	79.1	75.9	60.9	33.0	75.1
Contrast2Divide [30]	Best	96.4	95.3	**94.4**	**93.6**	**93.5**	78.7	76.4	**67.8**	**58.7**	**75.5**
	Last	96.2	95.2	94.3	93.4	90.8	78.3	76.1	67.4	58.5	75.1
CCLM (C2D)	Best	**96.5**	**95.4**	94.4	93.6	92.2	**79.4**	**76.9**	67.6	55.2	75.5
	Last	96.4	95.4	94.1	93.2	90.8	79.2	76.6	67.1	55.0	73.2

Table 2
Real noise performance comparison (test accuracy (%) on CIFAR-N datasets under different noise types). Bold indicates best performance. *Baseline results are reproduced*.

Method		CIFAR-10N					CIFAR-100N
		Aggregate	Random 1	Random 2	Random 3	Worst	Noisy fine
DivideMix [26]	Best	95.3	95.6	**95.6**	95.6	**93.2**	69.4
	Last	95.3	95.6	95.0	95.2	92.9	68.8
CCLM (DM)	Best	**95.6**	**95.6**	95.4	**95.6**	93.0	**70.7**
	Last	95.4	95.5	95.2	95.4	92.9	69.9
Contrast2Divide [30]	Best	95.7	**95.9**	**95.8**	93.8	**92.8**	**70.8**
	Last	95.6	95.6	95.8	93.0	92.0	70.3
CCLM (C2D)	Best	**95.8**	95.7	95.7	**95.9**	92.4	70.6
	Last	95.5	95.6	95.3	95.7	92.0	69.7

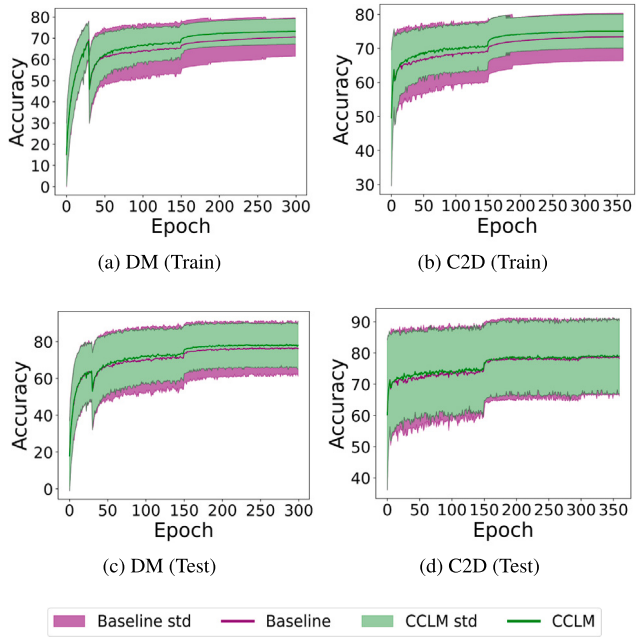

Fig. 4. Model accuracy and Classiness on CIFAR-100 20% Sym. noise. The accuracy drop is a result of the ending of the warm-up phase (at epoch 30 for DivideMix and 5 for C2D).

where $\mathrm{Acc}(m, D^j)$ is the accuracy of m on the jth class, N is the total number of samples and N^j is the number of samples in the jth class. Moreover, the term $\frac{N^j}{N}$ is a weight that extends the definition of classiness to class-imbalanced datasets. Classiness measures the spread of the model accuracy over the classes. Hence, low classiness is a desired property.

The evolution of model accuracy and classiness using 20% sym. noise of CIFAR-100 is shown in Fig. 4. The green and magenta lines show the accuracy of our method and of the concurrent method, respectively. For our method (concurrent method) the classiness is depicted as a green (magenta) area around the accuracy. We can observe that the training classiness is consistently smaller for the proposed CCLM (Figs. 4(a) and 4(b)). This shows that CCLM is less class-biased during training, yielding a better learning outcome. Finally, Table 3 shows the average and the best training and testing classiness on CIFAR-100 for the different noise levels, where the *best* epoch is chosen as the one with the best test accuracy. We differentiate between train and test classiness scores as the train classiness measures how much noise is being learned, and the test classiness measures the spread of the model accuracy on unseen data. It can be seen that CCLM has lower classiness in low-noise settings and comparable classiness in high-noise settings for both training and testing.

Table 3
Model classiness (Eq. (3)) on CIFAR-100. *Best* reports the classiness of the model at the epoch with the highest test accuracy. *Average* reports the average classiness of the model across all the epochs.

Noise			Method			
			DivideMix		CCLM (DM)	
			Average	Best	Average	Best
Train	Sym.	20%	11.0	9.0	7.7	**6.0**
		50%	14.3	12.7	10.9	**8.8**
		80%	18.2	17.6	17.5	**17.1**
		90%	18.1	18.8	17.1	**17.1**
	Asym. 40%		17.3	14.9	14.1	**11.2**
Test	Sym.	20%	16.0	14.0	13.6	**12.0**
		50%	16.7	15.0	14.9	**13.0**
		80%	20.0	19.0	19.4	**19.0**
		90%	19.8	**21.0**	20.2	22.0
	Asym. 40%		18.7	17.5	17.2	**14.0**

Noise division accuracy. To evaluate our label noise model, we check the performance on the Clean-Noisy split, at the first epoch after warm-up. We show the average (over two runs) division accuracy and the class standard deviation in Fig. 5 on CIFAR-10 and CIFAR-100 for

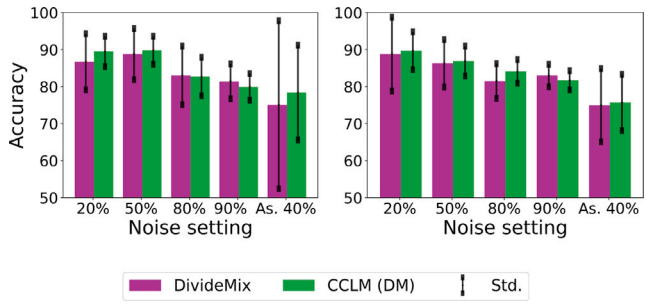

Fig. 5. Clean/Noisy split accuracy on CIFAR-10 (left) and CIFAR-100 (right) for all the synthetic noise settings.

Table 4
Impact of clean/noise split threshold on final model accuracy using CIFAR-100 with Sym. 80% and 90% noise. Bold indicates best performance.

Noise	GMM thres. (τ)	0.5	0.6	0.7	0.8	0.9
80%	DivideMix	**60.2**	**59.7**	59.1	58.7	54.7
	CCLM (DM)	59.0	58.4	**60.6**	**61.1**	**61.0**
90%	DivideMix	30.3	**31.5**	31.6	27.6	27.6
	CCLM (DM)	**30.6**	31.1	**32.6**	**33.5**	**30.8**

different noise settings, right after the warm-up. Our proposed method achieves a better accuracy on low noise levels, and a comparable accuracy in high noise levels. More importantly, we show how our method consistently achieves a smaller standard deviation, which results in a reduced bias towards easier classes.

Effect of clean-noise split threshold. We have highlighted the importance of the Clean-Noise split thoroughly, especially focusing on how to model label noise more sensibly. However, the Clean-Noise partition is ultimately decided by a threshold parameter. Therefore, an essential component of the analysis involves evaluating the influence of the chosen threshold. To understand the effect of the split threshold, we vary the GMM Threshold (τ) in DivideMix-based experiments. The SSL pre-training of C2D enables the models to determine the noisy examples with high confidence, making this study less relevant. As shown in Table 4 on CIFAR-100 with 80% and 90% symmetric noise, we observe that the model accuracy is consistent through the different values of τ. Fig. 3 presents the feature space that we aim to split. We argue that a class-agnostic label noise modelling, with a high threshold, would include less clean samples, whereas, with a low threshold, would include more noisy samples. In contrast, CCLM is not susceptible to this pitfall and is capable of handling higher thresholds.

6. Conclusions and future directions

Learning with Noisy Labels plays a pivotal role in data-centric deep learning due to its wide applications. In this paper, we study the division of clean-noisy samples in detail. Typical global noise modelling relies on the assumption that data is independent and normally distributed. We show that such a treatment is not always reasonable, and propose Class-Conditional Local Noise Model (CCLM) that relaxes these assumptions. In addition, we propose a new metric, *classiness*, that is important both for its obvious desirability and as a tool to understand training dynamics. Our future work focuses on analysing the impact of these relaxed assumptions on other parts of the tested algorithms. Tracking *classiness* for other datasets and, in particular, for class-imbalanced datasets is also an interesting research avenue since it can potentially open the path for further understanding of the training dynamics with this type of dataset. Finally, studying how other algorithms could benefit from this new label noise model is also envisaged.

CRediT authorship contribution statement

Albert Tatjer: Investigation, Methodology, Visualization, Writing – original draft. **Bhalaji Nagarajan:** Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ricardo Marques:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – review & editing. **Petia Radeva:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was partially funded by the EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), CERCA Programme/Generalitat de Catalunya, and Grants PID2022-141566NB-I00 (IDEATE), PDC2022-133642-I00 (DeepFoodVol), and CNS2022-135480 (A-BMC) funded by MICIU/AEI/10.13039/501100011033, by FEDER (UE), and by European Union NextGenerationEU/PRTR. R. Marques acknowledges the support of the Serra Húnter Programme. B. Nagarajan acknowledges the support of FPI Becas, MICINN, Spain.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2024.04.012.

References

[1] Y.-H. Liao, A. Kar, S. Fidler, Towards good practices for efficiently annotating large-scale image classification datasets, in: CVPR, 2021, pp. 4350–4359.
[2] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: A survey, Knowl.-Based Syst. 215 (2021) 106771.
[3] S. Li, X. Xia, J. Deng, S. Ge, T. Liu, Transferring annotator-and instance-dependent transition matrix for learning from crowds, 2023, arXiv preprint arXiv:2306.03116.
[4] C. Tan, J. Xia, L. Wu, S.Z. Li, Co-learning: Learning from noisy labels with self-supervision, in: ACM ICMM, 2021, pp. 1405–1413.
[5] C. Northcutt, L. Jiang, I. Chuang, Confident learning: Estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2021) 1373–1411.
[6] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, Y. Liu, Learning with noisy labels revisited: A study using real-world human annotations, in: ICLR, 2021.
[7] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., A closer look at memorization in deep networks, in: ICML, 2017, pp. 233–242.
[8] Z. Allen-Zhu, Y. Li, Y. Liang, Learning and generalization in overparameterized neural networks, going beyond two layers, Neural Inf. Process. Syst. 32 (2019).
[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.
[10] D. Angluin, P. Laird, Learning from noisy examples, Mach. Learn. 2 (1988) 343–370.
[11] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[12] Y. Zhang, G. Niu, M. Sugiyama, Learning noise transition matrix from only noisy labels via total variation regularization, in: ICML, PMLR, 2021, pp. 12501–12512.
[13] S. Wu, X. Xia, T. Liu, B. Han, M. Gong, N. Wang, H. Liu, G. Niu, Class2simi: A noise reduction perspective on learning with noisy labels, in: ICML, PMLR, 2021, pp. 11285–11295.

[14] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, Neural Inf. Process. Syst. 31 (2018).

[15] Y. Kim, J. Yun, H. Shon, J. Kim, Joint negative and positive learning for noisy labels, in: CVPR, 2021, pp. 9442–9451.

[16] A.K. Menon, A.S. Rawat, S.J. Reddi, S. Kumar, Can gradient clipping mitigate label noise? in: ICLR, 2020.

[17] Y. Chen, S.X. Hu, X. Shen, C. Ai, J.A. Suykens, Compressing features for learning with noisy labels, IEEE Trans. NNLS (2022).

[18] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: ICML, 2018, pp. 4334–4343.

[19] M. Xu, Z. Lian, L. Feng, B. Liu, J. Tao, DALI:Dynamically adjusted label importance for noisy partial label learning, 2023, arXiv:2301.12077.

[20] D. Ortego, E. Arazo, P. Albert, N.E. O'Connor, K. McGuinness, Multi-objective interpolation training for robustness to label noise, in: Proceedings of the IEEE/CVF CVPR, 2021, pp. 6606–6615.

[21] X. Xia, J. Deng, W. Bao, Y. Du, B. Han, S. Shan, T. Liu, Holistic label correction for noisy multi-label classification, in: ICCV, 2023, pp. 1483–1493.

[22] M. Wu, Q. Li, F. Yang, J. Zhang, V.S. Sheng, J. Hou, Learning from biased crowdsourced labeling with deep clustering, Expert Syst. Appl. 211 (2023) 118608.

[23] Y. Chen, X. Shen, Y. Liu, Q. Tao, J.A. Suykens, Jigsaw-ViT: Learning Jigsaw puzzles in vision transformer, Pattern Recognit. Lett. 166 (2023) 53–60.

[24] M. Zhang, N. Xu, X. Geng, Feature-induced label distribution for learning with noisy labels, Pattern Recognit. Lett. 155 (2022) 107–113.

[25] E. Arazo, D. Ortego, P. Albert, N. O'Connor, McGuinness, Unsupervised label noise modeling and loss correction, in: ICML, 2019, pp. 312–321.

[26] J. Li, R. Socher, S.C. Hoi, Dividemix: Learning with noisy labels as semi-supervised learning, 2020, arXiv preprint arXiv:2002.07394.

[27] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, L. Fei-Fei, Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, in: ICML, 2018, pp. 2304–2313.

[28] Y. Yao, Z. Sun, C. Zhang, F. Shen, Q. Wu, J. Zhang, Z. Tang, Jo-src: A contrastive approach for combating noisy labels, in: Proceedings of the IEEE/CVF CVPR, 2021, pp. 5192–5201.

[29] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C.A. Raffel, Mixmatch: A holistic approach to semi-supervised learning, Adv. Neural Inf. Process. Syst. 32 (2019).

[30] E. Zheltonozhskii, C. Baskin, A. Mendelson, A.M. Bronstein, O. Litany, Contrast to divide: Self-supervised pre-training for learning with noisy labels, in: WACV, 2022, pp. 1657–1667.

[31] B. Nagarajan, R. Marques, M. Mejia, P. Radeva, Class-conditional importance weighting for deep learning with noisy labels, in: VISIGRAPP (5: VISAPP), 2022, pp. 679–686.

[32] A. Tatjer, B. Nagarajan, R. Marques, P. Radeva, CCLM: class-conditional label noise modelling, in: IBPRIA, Springer, 2023, pp. 3–14.

[33] Y. Bai, E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu, T. Liu, Understanding and improving early stopping for learning with noisy labels, Adv. Neural Inf. Process. Syst. 34 (2021) 24392–24403.

[34] Y. Tu, B. Zhang, Y. Li, L. Liu, J. Li, J. Zhang, Y. Wang, C. Wang, C.R. Zhao, Learning with noisy labels via self-supervised adversarial noisy masking, in: Proceedings of the IEEE/CVF CVPR, 2023, pp. 16186–16195.

[35] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, J. Bailey, Symmetric cross entropy for robust learning with noisy labels, in: ICCV, 2019, pp. 322–330.

[36] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, J. Bailey, Normalized loss functions for deep learning with noisy labels, in: ICML, 2020, pp. 6543–6553.

[37] Z. Liu, X. Zhang, J. He, D. Fu, D. Samaras, R. Tan, X. Wang, S. Wang, ChiMera: Learning with noisy labels by contrasting mixed-up augmentations, 2023, arXiv preprint arXiv:2310.05183.

[38] N. Karim, M.N. Rizve, N. Rahnavard, A. Mian, M. Shah, Unicon: Combating label noise through uniform selection and contrastive learning, in: Proceedings of the IEEE/CVF CVPR, 2022, pp. 9676–9686.

[39] Z. Huang, J. Zhang, H. Shan, Twin contrastive learning with noisy labels, in: Proceedings of the IEEE/CVF CVPR, 2023, pp. 11661–11670.

[40] E. Kazemi, F. Taherkhani, L. Wang, On complementing unsupervised learning with uncertainty quantification, Pattern Recognit. Lett. 176 (2023) 69–75.

[41] K. Wang, C. Zhang, Y. Geng, H. Ma, Evidential pseudo-label ensemble for semi-supervised classification, Pattern Recognit. Lett. 177 (2024) 135–141.

[42] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, M. Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, Neural Inf. Process. Syst. 31 (2018).

[43] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, M. Sugiyama, How does disagreement help generalization against label corruption? in: ICML, PMLR, 2019, pp. 7164–7173.

[44] H. Wei, L. Feng, X. Chen, B. An, Combating noisy labels by agreement: A joint training method with co-regularization, in: Proceedings of the IEEE/CVF CVPR, 2020, pp. 13726–13735.

[45] K. Cai, H. Zhang, W. Pedrycz, D. Miao, SSS-net: A shadowed-sets-based semi-supervised sample selection network for classification on noise labeled images, Knowl.-Based Syst. (2023) 110732.

[46] Y. Li, H. Han, S. Shan, X. Chen, DISC: Learning from noisy labels via dynamic instance-specific selection and correction, in: CVPR, 2023, pp. 24070–24079.

[47] X. Yu, Y. Jiang, T. Shi, Z. Feng, Y. Wang, M. Song, L. Sun, How to prevent the continuous damage of noises to model training? in: Proceedings of the IEEE/CVF CVPR, 2023, pp. 12054–12063.

[48] X. Xia, B. Han, Y. Zhan, J. Yu, M. Gong, C. Gong, T. Liu, Combating noisy labels with sample selection by mining high-discrepancy examples, in: Proceedings of the IEEE/CVF ICCV, 2023, pp. 1833–1843.

[49] C. Ahn, K. Kim, J. Baek, J. Lim, Han, Sample-wise label confidence incorporation for learning with noisy labels, in: ICCV, 2023, pp. 1823–1832.

[50] K. Nishi, Y. Ding, A. Rich, T. Hollerer, Augmentation strategies for learning with noisy labels, in: CVPR, 2021, pp. 8022–8031.

[51] H. Wang, R. Xiao, Y. Dong, L. Feng, J. Zhao, ProMix: Combating label noise via maximizing clean sample utility, 2022, arXiv:2207.10276.

[52] D. Kim, K. Ryoo, H. Cho, S. Kim, SplitNet: Learnable clean-noisy label splitting for learning with noisy labels, 2022, arXiv:2211.11753.

[53] Z. Sun, F. Shen, D. Huang, Q. Wang, X. Shu, Y. Yao, J. Tang, Pnp: Robust learning from noisy labels by probabilistic noise prediction, in: Proceedings of the IEEE/CVF CVPR, 2022, pp. 5311–5320.

[54] F. Fooladgar, M.N.N. To, P. Mousavi, P. Abolmaesumi, Manifold DivideMix: A semi-supervised contrastive learning framework for severe label noise, 2023, arXiv preprint arXiv:2308.06861.

[55] B. Nagarajan, R. Marques, E. Aguilar, P. Radeva, Bayesian DivideMix++ for enhanced learning with noisy labels, Neural Netw. (2024) 106122.

[56] Y. Huang, B. Bai, S. Zhao, K. Bai, F. Wang, Uncertainty-aware learning against label noise on imbalanced datasets, in: ICAI, 2022, pp. 6960–6969.

[57] J. Shu, X. Yuan, D. Meng, Z. Xu, Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[58] J. Huang, Y. Chen, J. Feng, X. Wu, Class prototype-based cleaner for label noise learning, 2022, arXiv preprint arXiv:2212.10766.

[59] J. Yao, B. Han, Z. Zhou, Y. Zhang, I.W. Tsang, Latent class-conditional noise model, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[60] Q. Wei, L. Feng, H. Sun, R. Wang, C. Guo, Y. Yin, Fine-grained classification with noisy labels, in: CVPR, 2023, pp. 11651–11660.

[61] G. Valle-Perez, C.Q. Camargo, A.A. Louis, Deep learning generalizes because the parameter-function map is biased towards simple functions, 2018, arXiv preprint arXiv:1805.08522.

[62] A. Krizhevsky, G. Hinton, et al., Learning Multiple Layers of Features from Tiny Images, Toronto, ON, Canada, 2009.

[63] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE CVPR, 2017, pp. 1944–1952.

[64] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 630–645.

[65] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.

[66] D. Tanaka, D. Ikami, T. Yamasaki, K. Aizawa, Joint optimization framework for learning with noisy labels, in: Proceedings of the IEEE CVPR, 2018, pp. 5552–5560.

[67] K. Yi, J. Wu, Probabilistic end-to-end noise correction for learning with noisy labels, in: Proceedings of the IEEE/CVF CVPR, 2019, pp. 7017–7025.