Mathematical Biosciences and Engineering

MBE, 20(10): 17905–17918.
DOI: 10.3934/mbe.2023796
Received: 08 July 2023
Revised: 02 September 2023
Accepted: 12 September 2023
Published: 19 September 2023

http://www.aimspress.com/journal/MBE

Research article

Complementary label learning based on knowledge distillation

Peng Ying, Zhongnian Li, Renke Sun and Xinzheng Xu*

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

* **Correspondence:** Email: xxzheng@cumt.edu.cn; Tel: +8615952151616; Fax: +8651683591726.

Abstract: Complementary label learning (CLL) is a type of weakly supervised learning method that utilizes the category of samples that do not belong to a certain class to learn their true category. However, current CLL methods mainly rely on rewriting classification losses without fully leveraging the supervisory information in complementary labels. Therefore, enhancing the supervised information in complementary labels is a promising approach to improve the performance of CLL. In this paper, we propose a novel framework called Complementary Label Enhancement based on Knowledge Distillation (KDCL) to address the lack of attention given to complementary labels. KDCL consists of two deep neural networks: a teacher model and a student model. The teacher model focuses on softening complementary labels to enrich the supervision information in them, while the student model learns from the complementary labels that have been softened by the teacher model. Both the teacher and student models are trained on the dataset that contains only complementary labels. To evaluate the effectiveness of KDCL, we conducted experiments on four datasets, namely MNIST, F-MNIST, K-MNIST and CIFAR-10, using two sets of teacher-student models (Lenet-5+MLP and DenseNet-121+ResNet-18) and three CLL algorithms (PC, FWD and SCL-NL). Our experimental results demonstrate that models optimized by KDCL outperform those trained only with complementary labels in terms of accuracy.

Keywords: weakly supervised learning; complementary label learning; knowledge distillation; deep neural networks; deep learning

1. Introduction

Supervised learning is an important branch of machine learning. In supervised multi-classification

problems, each sample is assigned a label which indicates the category it belongs to [1]. Supervised learning is effective when there are enough samples with high quality labels. However, it is expensive and time-consuming to build datasets with a multitude of accurate labels. To solve this problem, researchers have proposed a series of weakly supervised learning (WSL) methods, which aim to train models with partial, incomplete or inaccurate supervised information, such as noise-label learning [2–5], semi-supervised learning [6–9], partial-label learning [10–12], positive-confidence learning [13], unlabeled-unlabeled learning [14] and others.

In this paper, we consider another WSLframework called complementary label learning (CLL). We show the difference between complemtary labels and true labels in Figure 1. Compared to an ordinary label, a complementary label indicates the class that the sample does not belong to. Obviously, it is easier and less costly to collect these complementary labels. For example, in some very specialized domains, the expert knowledge is very expensive. If complementary labels are used for annotation, we need to only determine the extent of the label space and then use common sense to determine which category is wrong. It is much simpler and faster to determine which class a sample does not belong to than it belongs to. Besides, CLL can also protect data privacy in some sensitive fields like medical and financial records because we no longer need to disclose the true information of the data. This not only protects data privacy and security, but also makes it easier to collect data in these areas.

The framework of CLL was first proposed by Ishida et al. [15]. They proved that the unbiased risk estimator (URE) only from complementary labels is equivalent to the ordinary classification risk when the loss function satisfies certain conditions. In URE, the loss function must be nonconvex and symmetric which leads to certain limitations. To overcome this limitation, Yu et al. [16] made cross-entropy loss usable in CLL by constructing a complementary label transition matrix, and they also considered that different labels had different probability of being selected as a complementary label. Then, Ishida et al. [17] expanded URE and proposed a CLL framework adapted to more general loss functions. This framework still has an unbiased estimator of the regular classification risk, but it works for all loss functions. Chou et al. [18] optimized URE from gradient estimation, and proposed that using surrogate complementary loss (SCL) to obtain unbiased risk estimation, which effectively alleviated the problem of overfitting in URE. Liu et al. [19] applied common losses such as categorical cross entropy (CCE), mean square error (MSE) and mean absolute error (MAE) to CLL. Ishiguro et al. [20] conducted a study on the problem that complementary labels may be affected by label noise. To mitigate its adverse effects, they selected losses with noise robustness which satisfied weighted symmetric condition or a more relaxed condition. Recently, Zhang et al. [21] broadened the setting of complementary label datasets and discussed the case that the datasets contained a large number of complementary labels and a small number of true labels at the same time. They proposed an adversarial complementary label learning network, named Clarinet. Clarinet consists of two deep neural networks, one to classify complementary labels and true labels, and the other to learn from complementary labels.

Previous studies on CLL always focus on rewriting the classification risk under the ordinary label distribution to the risk under the complementary label distribution and exploring the use of more loss functions [15–19]. These rewriting risk techniques prove the consistency relationship between the risk of complementary label classification and the risk of supervised classification. This enables the classifier to perform accurate classification using only the complementary labels. However, in this process, only complementary labels are involved in the risk calculation, and the information contained in them is extremely limited, which results in consistently lower performance of CLL compared to supervised learning. Therefore, we aim to enhance the supervision information of the complementary

labels to further improve the performance of CLL. In this paper, we propose a two-step complementary label enhancement framework based on knowledge distillation (KDCL). It consists of the following components: 1) a teacher model trained on complementary label dataset to generate soft labels which contain more supervision information as label distribution; 2) a student model trained on the same dataset to learn from both soft labels and complementary labels; 3) a final loss function to integrate loss from soft labels and complementary labels and update parameters of the student model. We use three CLL loss functions to conduct experiments on several benchmark datasets, and compare the accuracy of the student model before and after enhancement by KDCL. The experimental results show that KDCL can effectively improve the performance of CLL.

Figure 1. Comparison of the complementary labels (bottom) with the real labels (top). Complementary label is one of categories the image does not belong to.

2. Preliminaries

2.1. Learning from true labels

Supposing that the input sample is a d-dimensional vector $x \in \mathbb{R}^d$ with class labels $y \in \{1,2,\ldots,K\}$, where K stands for K classes in the dataset. Giving a training set $D = \{(x_i, y_i)\}_{i=1}^N$ with N samples, all of which independently follow the same distribution $p(x, y)$. The goal of learning from true labels is to learn a mapping relation $f(x)$ from the sample space $\mathbb{R}^d$ to the label space $\{1,2,\ldots,K\}$ and $f(x)$ is also called a classifier. We want $f(x)$ to minimize the multi-class classification risk:

$$R(f) = \mathbb{E}_{p(x,y)\sim D}[L(f(x), y)], \tag{1}$$

where $L(f(x), y)$ is multi-class loss function, $f(x)$ is usually obtained by the following equation:

$$f(x) = argmax_{y\in 1,2,\ldots,K} g_y(x), \tag{2}$$

where $g(x)\colon \mathbb{R}^d \to \mathbb{R}^K$. In deep neural networks, $g(x)$ is the prediction distribution of the output from the last fully connected layer.

In general, distribution $p(x, y)$ is unknown. We can use the sample mean to approximate the classification risk in Eq (1). $R(f)$ is empirically estimated as $\hat{R}(f)$:

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{n} L(f(x_i), y_i), \tag{3}$$

where N is the number of training data and i is the i-th sample.

2.2. Learning from complementary labels

In CLL, each sample x is assigned only one complementary label $\bar{y}$. Therefore, the dataset is switched from $D = \{(x_i, y_i)\}_{i=1}^{N}$ to $\bar{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{N}$, where $\bar{y} \in \{1,2,\ldots,K\}\backslash\{y\}$ and $D \neq \bar{D}$. $\bar{D}$ independently follow an unknown distribution $\bar{p}(x, \bar{y})$. If all complementary labels are selected in an unbiased way, which means that they have the same probability of being chosen, $\bar{p}(x, \bar{y})$ can be presented as:

$$\bar{p}(x, \bar{y}) = \frac{1}{K-1} \sum_{y \neq \bar{y}} p(x, y). \tag{4}$$

Supposing that $\bar{L}(f(x), \bar{y})$ is complementary loss function, we can obtain similar multi-class risk as Eq (1) in distribution $\bar{p}(x, \bar{y})$:

$$\bar{R}(f) = \mathbb{E}_{\bar{p}(x,\bar{y})\sim\bar{D}}[\bar{L}(f(x), \bar{y})]. \tag{5}$$

To our best knowledge, Ishida et al. [15] are the first to prove that the difference between Eq (1) and Eq (5) is constant when the loss function $\bar{L}$ satisfies certain conditions and this constant M only depends on the number of categories K:

$$\begin{aligned} R(f) &= (K-1)\mathbb{E}_{\bar{p}(x,\bar{y})\sim\bar{D}}[\bar{L}(f(x), \bar{y})] + M \\ &= (K-1)\bar{R}(f) + M. \end{aligned} \tag{6}$$

All coefficients are constant when the loss function satisfies the condition. So it is possible to learn from complementary labels by minimizing $R(f)$ in Eq (6). Then, they rewrite one-versus-all (OVA) loss L_{OVA} and pairwise-comparison (PC) loss L_{PC} in ordinary multi-class classification as $\bar{L}_{OVA}$ and $\bar{L}_{PC}$ in CLL:

$$\begin{aligned} \bar{L}_{OVA}(g(x), \bar{y}) &= \frac{1}{K-1} \sum_{y \neq \bar{y}} l\left(g_y(x)\right) + l\left(-g_{\bar{y}}(x)\right), \\ \bar{L}_{PC}(g(x), \bar{y}) &= \sum_{y \neq \bar{y}} l\left(g_y(x) - g_{\bar{y}}(x)\right), \end{aligned} \tag{7}$$

where $l(z): \mathbb{R} \to \mathbb{R}$ is a binary loss and it must be nonconvex and symmetric, such as sigmoid loss. $g(x)$ is the same as Eq (2) and $g_y(x)$ is the y-th element of $g(x)$. Finally, the unbiased risk estimator of $R(f)$ can be obtained by sample mean:

$$\hat{R}(f) = \frac{(K-1)}{N} \sum_{n=1}^{N} \bar{L}(f(x_n), \bar{y}_n) + M. \tag{8}$$

Although it is feasible to learn a classifier that minimizes Eq (8) from complementary labels, the restriction on the loss function limits the application of URE. Yu et al. [16] analyze the relationship between ordinary and complementary labels in terms of conditional probability:

$$P(\bar{y} = j|x) = \sum_{i \neq j} P(\bar{y} = j|y = i)\, P(y = i|x), \tag{9}$$

where $\forall i, j \in \{1,2, \ldots, K\}$. When all complementary labels are selected in an unbiased way, $P(\bar{y}|y)$ can be expressed as a transition matrix Q:

$$Q = \begin{bmatrix} 0 & \cdots & \frac{1}{K-1} \\ \vdots & \ddots & \vdots \\ \frac{1}{K-1} & \frac{1}{K-1} & 0 \end{bmatrix}_{K \times K}, \tag{10}$$

where each element in Q represents $P(\bar{y} = j|y = i)$. Since the true label and the complementary label of the sample are mutually-exclusive, that is $P(\bar{y} = j|y = i) = 0$. Therefore, the entries on the diagonal of the matrix are 0.

Combining Eqs (5), (9) and (10), we can rewrite $\bar{R}(f)$ as:

$$\bar{R}(f) = \mathbb{E}_{\bar{p}(x,\bar{y})}[L_{CE}(Q^T g(x), \bar{y})], \tag{11}$$

where L_{CE} is cross-entropy loss which is widely used in deep learning. The classification risk $\bar{R}(f)$ in Eq (8) is also consistent with the ordinary classification risk $R(f)$ [16].

3. Complementary label learning based on knowledge distillation

3.1. Framework architecture

In image classification, outputs from the last fully connected layer of a deep neural network contain the predicted probability distribution of all classes after the Softmax function. Comparing with a single logical label, the outputs carry more information. Hinton et al. [22] define the outputs as soft labels and propose a knowledge distillation framework. We draw on the idea of knowledge distillation and hope to improve the performance of CLL by enhancing complementary labels through soft labels.

In the framework of knowledge distillation, Hinton et al. [22] modify the Softmax function and they introduce the parameter T to control the smoothness of soft labels. The ordinary Softmax function can be expressed as follows:

$$y_i' = \frac{exp(y_i)}{\sum_j exp(y_j)}, \tag{12}$$

where y_i' is the predicted probability of the i-th class, $exp(\cdot)$ is the exponential function and y_i is the predicted output of the classification network for the ith class. The Softmax function combines the prediction outputs of the model for all classes, and uses the exponential function to normalize the

output values in the interval [0,1].

The rewritten Softmax function is as follows:

$$y_i' = \frac{exp(^{y_i}/_T)}{\sum_j exp\left(^{y_j}/_T\right)}. \tag{13}$$

We present a comparison of the smoothness of soft labels for different T in Figure 2. As T gradually increases, soft labels will become smoother. Actually, T regulates the degree to the attention to the negative labels. The higher T, the more attention is paid to negative labels. T is an adjustable hyperparameter during training.

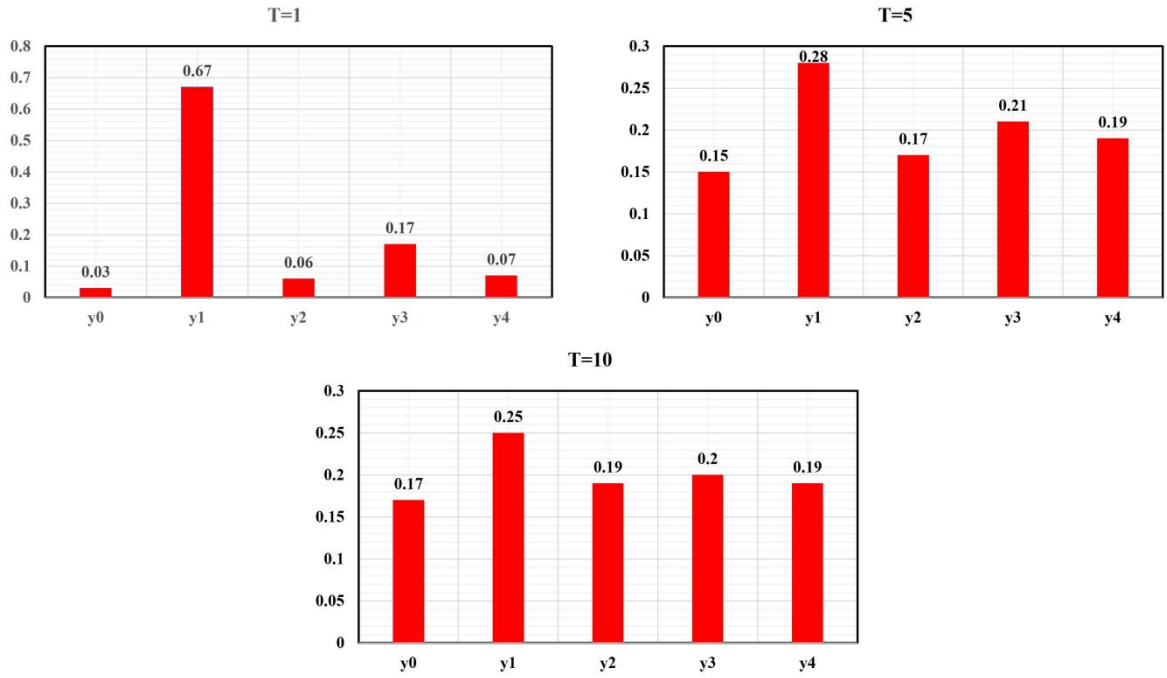


Figure 2. The smoothness of soft labels for different T. The higher T, the smoother soft labels will be.

For one sample, soft labels not only clarify its correct category, but also contain the correlation between other labels. More abundant information is carried in soft labels than the complementary label. If we add an extra term to the ordinary supplementary label classification loss and introduce soft labels as additional supervision information, CLL will perform better than using only complementary labels. Of course, we need a model with high accuracy to produce soft labels, which will make the soft labels more credible. This model is also trained by complementary labels.

Taking advantage of this property, we propose KDCL, a complementary label learning framework based on knowledge distillation. The overall structure is shown in Figure 3.

KDCL is a two-stage training framework consisting of a more complex teacher model with higher accuracy and a simpler student model with lower accuracy. First, the teacher model is trained with complementary labels on the dataset and predicts all samples in the training set. The prediction results are normalized by the Softmax function with $T = t(t > 1)$ to generate soft labels S_{tea}. Second, the student model is trained and its outputs are processed in two ways, one to produce the soft prediction

results S_{stu} with $T = t(t > 1)$, and the other to output ordinary prediction results P_{stu} with $T = 1$. Then, the KL divergence between S_{tea} and S_{stu} is calculated, and the complementary label loss between P_{stu} and the complementary labels is calculated at the same time. The two losses are weighted to obtain the final distillation loss. Finally, parameters of the student model will be updated by the final loss.

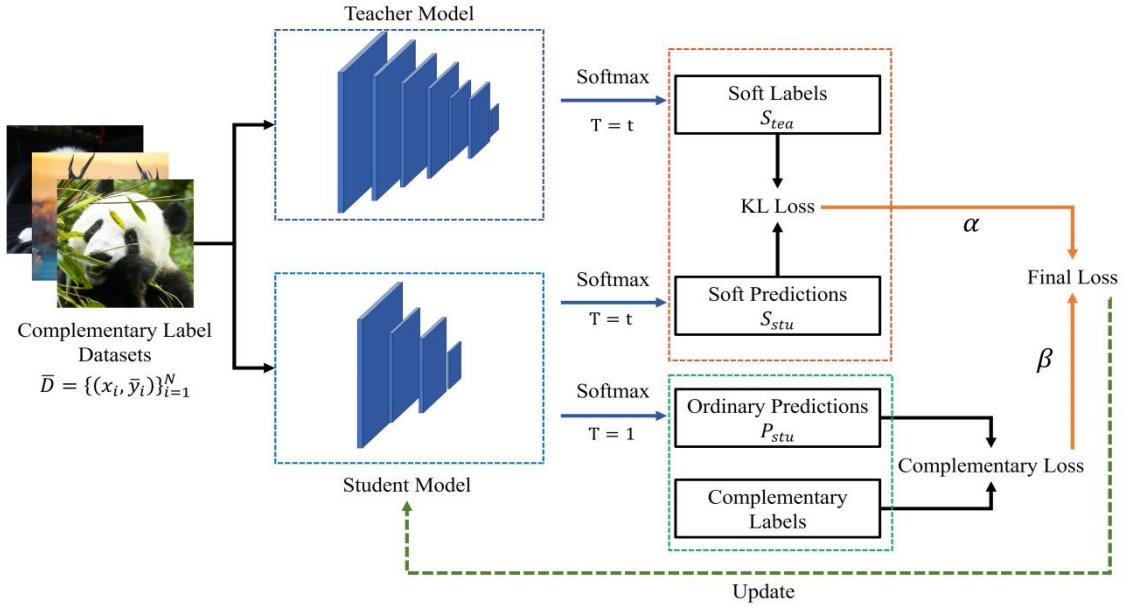


Figure 3. The framework architecture of KDCL. α and β are the weighting factors to balance KL loss and complementary loss.

In KDCL, the final loss consists of Kullback-Leible (KL) loss and complementary loss. On the one hand, the student model needs to learn knowledge from the teacher model to improve its ability. On the other hand, the teacher model is not completely correct, and the student model also needs to learn by itself to reduce the influence of the teacher model's errors on the learning process. It is better to consider both of them.

3.2. Loss function design

The final distillation loss consists of two parts and it can be expressed as follows:

$$L_{KDCL} = \alpha L_{KL} + L_{CL}, \tag{14}$$

where L_{KL} denotes the KL divergence and L_{CL} denotes the complementary loss. Given the probability distributions p_t from the teacher model and p_s from the student model, their KL divergence can be expressed as follows:

$$L_{KL}(p_t, p_s) = \sum_i -p_{ti} \, log \frac{p_{si}}{p_{ti}}, \tag{15}$$

where i denotes the i-th element in tensor p_t or p_s.

We select three complementary losses for KDCL. They are the PC loss proposed by Ishida et al. [15],

FWD loss proposed by Yu et al. [16] and SCL-NL loss proposed by Chou et al. [18]. Supposing that p_s is the probability distribution for sample x from the student model and $\bar{y}$ is the complementary label of x, these complementary losses are shown in Eqs (16)–(18).

$$\bar{L}_{PC}(p_s, \bar{y}) = \frac{K-1}{n} \sum_{y \neq \bar{y}} \left(p_{sy} - p_{s_{\bar{y}}} \right) - \frac{K \times (K-1)}{2} + K - 1, \tag{16}$$

$$\bar{L}_{FWD}(p_s, \bar{y}) = -\sum_i \bar{y}_i \times log(Q^T \times p_{s_i}), \tag{17}$$

$$\bar{L}_{SCL-NL}(p_s, \bar{y}) = \sum_i \bar{y}_i \times \left(-log\left(1 - p_{s_{\bar{y}}}\right) \right), \tag{18}$$

where K denotes the number of categories of the dataset, and Q^T denotes the transpose of Q which is a $K \times K$ square matrix with all entries $1/(K-1)$ except the diagonal.

With parameters p_t, p_s and $\bar{y}$, the final loss can be expressed in more detail as follows:

$$L_{KD-PC}(p_t, p_s, \bar{y}) = \alpha L_{KL}(p_t, p_s) + \bar{L}_{PC}(p_s, \bar{y}) \tag{19}$$

$$L_{KD-FWD}(p_t, p_s, \bar{y}) = \alpha L_{KL}(p_t, p_s) + \bar{L}_{FWD}(p_s, \bar{y}) \tag{20}$$

$$L_{KD-SCL}(p_t, p_s, \bar{y}) = \alpha L_{KL}(p_t, p_s) + \bar{L}_{SCL-NL}(p_s, \bar{y}) \tag{21}$$

α is the weighting factor, which is used to control the degree of influence of soft labels on the overall classification loss. The values of α will be determined in the experiment.

4. Experiments

We evaluate and compare the student models optimized by KDCL with the same models only trained by complementary labels on four public image classification datasets. Three complementary label losses including PC loss [15], FWD loss [16] and SCL-NL loss [18], are used as loss functions for training the models. All the experiments are carried out on a server with a 15 vCPU Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz, 80 GB RAM and one RTX 3090 GPU with 24 GB memory.

4.1. Datasets

Four benchmark image classification datasets, including MNIST, Fashion-MNIST(F-MNIST), Kuzushiji-MNIST(K-MNIST) and CIFAR10, are used to verify the effectiveness of KDCL.

MNIST: consists of 60,000 28 × 28 pixel grayscale images for training and 10,000 images for testing, with a total of 10 categories representing numbers between 0 and 9.

F-MNIST: is an alternative dataset to MNIST and consists of 10 categories, 60,000 training images and 10,000 test images, each with a size of 28 × 28 pixels.

K-MNIST: is a dataset derived from 10 Japanese ancient characters widely used between the mid-Heian period and early modern Japan, which is an extension of the MNIST dataset. K-MNIST contains a total of 74,000 gray-scale images of 28 × 28 pixels in 10 categories.

CIFAR10: consists of 60,000 32 × 32 color images, 50,000 of which are used as the training set and 10,000 as the test set. Each category contains 6000 images.

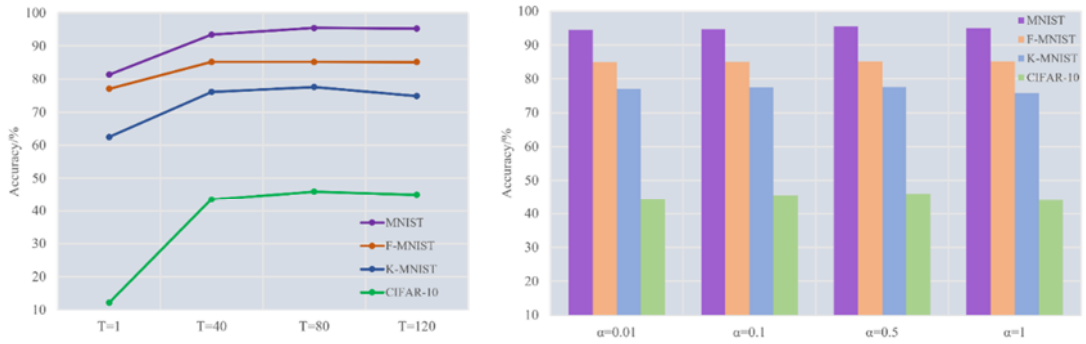
4.2. Experimental settings

Following the settings in [15,17,18], we use an unbiased way to select complementary labels for samples in all datasets. Besides, we apply two different sets of teacher-student networks to these datasets. Specifically, for MNIST, F-MNIST and K-MNIST, we chose Lenet-5 [23] as the teacher model and MLP [24] with 500 hidden neurons as the student model. Because these datasets are relatively simple, simple networks can work well. For CIFAR10 dataset, since color images are more difficult to be classified, we need deeper CNN to extract features. We choose DenseNet-121 [25] as the teacher model and ResNet-18 [26] as the student model.

In the setting of training details, for MNIST, F-MNIST and K-MNIST, we train Lenet-5 and MLP with 120 epochs and use SGD as the optimizer with a momentum 0.9 and a weight decay of 0.0001. The initial learning rate is 0.1 and it is halved every 30 epochs. The batch size is set to 128. For CIFAR10 dataset, we train DenseNet-121 and ResNet-18 with 80 epochs and use SGD as the optimizer with a momentum 0.9 and a weight decay of 0.0005. The learning rate is from {1e-1, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4} and it is divided by 10 every 30 epochs.

4.3. Parameter sensitivity analysis

In Figure 4, we make a parameter sensitivity analysis of the distillation temperature T in Eq (13) and the soft label weighting factor α in Eqs (19)–(21).



(a) Parametric analysis of T (b) Parametric analysis of α

Figure 4. Test accuracy results of different T with fixed α and comparison results of different α with fixed T. The experiments are conducted with Lenet-5 and MLP on MNIST, F-MNIST, K-MNIST and Desenet-121 and Resnet-18 on CIFAR-10.

We first explore the influence of different distillation temperature T. As we can see, when $T = 1$, which means directly using the probability distribution output by the teacher model as soft labels without softening, KDCL exhibits the worst accuracy. This is because when the temperature is low, there is a significant difference in soft labels between positive and negative classes, making it difficult for the student model to learn effectively. As T gradually increases, the soft labels become more and more smooth, and student model can easily learn the knowledge in soft labels, and the accuracy is

gradually improved. When $T \geq 80$, the gap between positive and negative classes in soft labels is extremely small, as well as the influence of negative classes is too large, which leads to the accuracy no longer increasing, or even decreasing.

Then, we further investigate the optimal value of soft label weighting factor α. We follow the setting in Hinton et al. [22], and set α in the range of 0 to 1. On the same dataset, the change of α does not have a great impact on the accuracy of KDCL. This indicates that the KDCL model parameter optimization process is not sensitive to the hyperparameter α. Nevertheless, the model still achieves higher accuracy when $\alpha = 0.5$.

Based on the above analysis, we will set $T = 80$, $\alpha = 0.5$ in subsequent experiments.

4.4. Experimental results

We show the accuracy for all models with three complementary label losses before and after being optimized by KDCL on four datasets. The results are presented in Table 1.

Table 1. Comparison of classification accuracies between different methods using different network architectures on MNIST, F-MNIST, K-MNIST and CIFAR-10.

Dataset	MNIST			F-MNIST			K-MNIST			CIFAR-10		
Model	Lenet-5	MLP	KDCL-MLP	Lenet-5	MLP	KDCL-MLP	Lenet-5	MLP	KDCL-MLP	Lenet-5	MLP	KDCL-MLP
PC	89.94%	83.78%	**86.10%**	77.22%	76.67%	**77.42%**	67.77%	**60.52%**	60.34%	38.31%	32.74%	**33.37%**
FWD	85.35%	83.67%	**84.61%**	85.35%	83.67%	**84.61%**	86.85%	70.86%	**75.41%**	60.74%	44.93%	**46.65%**
SCL-NL	98.18%	92.06%	**94.33%**	85.93%	83.69%	**84.66%**	86.85%	70.59%	**75.25%**	61.64%	40.46%	**45.98%**

In Table 1, we show the experimental results of KDCL, where we compare the performance of the student model optimized by KDCL with that trained only with complementary labels across different losses and datasets. On MNIST, which is a relatively simple and easy dataset, all methods can achieve high accuracies. With the help of KDCL, we improve the accuracy of MLP from 83.78% to 86.10% with PC loss, 92.07% to 94.32% with FWD loss and 92.06% to 94.33% with SCL-NL loss. SCL-NL loss performs better among three loss functions. Besides, after being enhanced by KDCL, the accuracy of KDCL-MLP falls between the accuracy of MLP model and Lenet-5. On F-MNIST, which is more complex than MNIST, all methods have a slight decrease. Our KDCL achives 77.42% with PC loss, 84.61% with FWD loss and 84.66% with SCL-NL loss. On K-MNIST, which is more complex than F-MNIST, when using PC loss, our method does not significantly improve the accuracy of MLP, but we improve 4.55% with FWD loss and 4.66% with SCL-NL loss. On CIFAR-10, which is the most complex among the four datasets, there is a significant drop in accuracies. Nevertheless, the student model can still be optimized by KDCL, demonstrating its robustness and effectiveness across different datasets.

We show the testing process of all models in Figure 5.

In Figure 5, we present the convergence speed of all models in our experiments. The results show that the student model distilled by KDCL converges faster than that trained only with complementary labels. This indicates that the model can learn the features of the images more accurately and efficiently when utilizing both soft labels and complementary labels.

Additionally, we observe that the PC loss exhibits a decrease in accuracy on more challenging

datasets, particularly on CIFAR10. This is because the PC loss uses the Sigmoid function as the normalization function, which can lead to negative values in the loss calculation and prevent the model from finding better parameters when updating. This phenomenon becomes more pronounced on the CIFAR10 dataset, where a peak appears. However, KDCL can alleviate this phenomenon and shift the peak to a later epoch. This demonstrates the effectiveness of KDCL in addressing the limitations of existing CLL methods and improving the performance of complementary label learning.

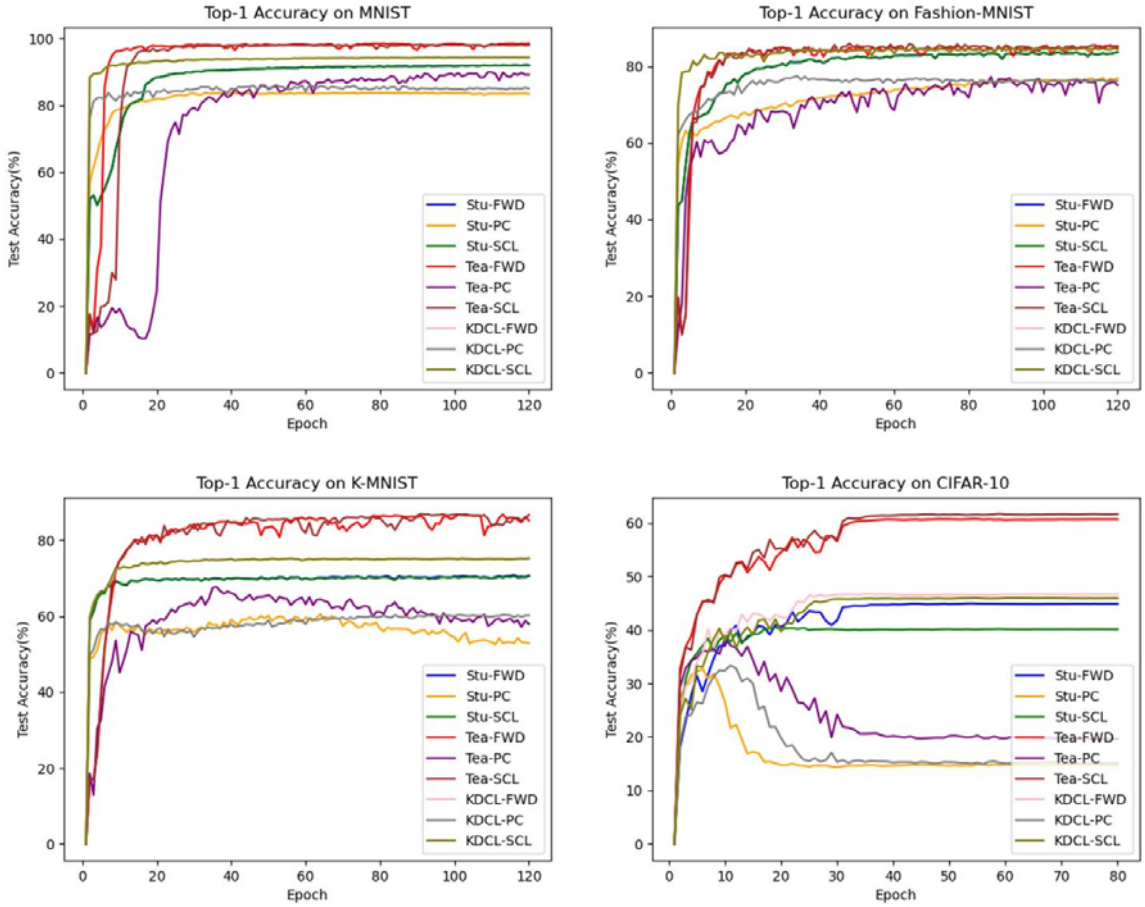


Figure 5. Comparison of the testing process of teacher models, student models and KDCL-student models on four datasets.

5. Discussion

In this study, we established a knowledge distillation training framework for CLL, called KDCL. As stated in the introduction, the supervision information in complementary labels is easily missed. The proposed framework employed a deep CNN model with higher accuracy to soften complementary labels to soft labels. Both soft labels and origion complementary labels are used to train the classification model. After the optimization of KDCL, compared to just using the normal CLL methods, the accuracy has been improved by 0.5–4.5%.

The main limitation lies in multiple aspects. First, KDCL's performance could be influenced by the choice of teacher-student models and CLL algorithms. Our experiments utilize specific

combinations of models and algorithms, and the results may vary with different configurations. By choosing better CNN networks and more excellent CLL algorithms, KDCL can achieve better performance on more difficult datasets. Another drawback of the proposed scheme is time cost. Due to the two-stage training framework of KDCL, which involves training a high-accuracy teacher model using complementary labels, the overall training time cost of KDCL is relatively high. Training a high-accuracy model typically takes a considerable amount of time, which poses a challenge to the efficiency of KDCL. In addition, KDCL is only tested on public datasets, and the data distribution is relatively uniform. In the future, we also consider expanding the application scope of KDCL to use dynamically imbalanced data for CLL, or to combine with hybrid deep learning models [27–29].

6. Conclusions

In this paper, we give the first attempt to leverage the knowledge distillation training framework in CLL. To enhance the supervised information present in complementary labels, which are often overlooked in existing CLL methods, we propose a complementary label enhancement framework based on knowledge distillation, called KDCL. Specifically, KDCL consists of a teacher model and a student model. By adopting knowledge distillation techniques, the teacher model transfers its softened knowledge to the student model. The student model then learns from both soft labels and complementary labels to improve its classification performance. The experimental results on four benchmark datasets show that KDCL can improve the classification accuracy of CLL, and maintain robustness and effectiveness on difficult datasets.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61976217, 62306320), the Natural Science Foundation of Jiangsu Province (No. BK20231063), the Fundamental Research Funds of Central Universities (No. 2019XKQYMS87), Science and Technology Planning Project of Xuzhou (No. KC21193).

Conflict of interest

All authors declare that they have no conflicts of interest.

References

1. Y. Katsura, M. Uchida, Bridging ordinary-label learning and complementary-label learning, in *Proceedings of the 12th Asian Conference on Machine Learning (ACML)*, **129** (2020), 161–176.
2. Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, L. J. Li, Learning from noisy labels with distillation, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, **97** (2017), 1928–1936. https://doi.org/10.1109/ICCV.2017.211

3. M. Hu, H. Han, S. Shan, X. Chen, Weakly Supervised image classification through noise regularization, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, (2019), 11509–11517. https://doi.org/10.1109/CVPR.2019.01178
4. K. H. Lee, X. He, L. Zhang, L. Yang, CleanNet: Transfer learning for scalable image classifier training with label noise, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, (2018), 5447–5456. https://doi.org/10.1109/CVPR.2018.00571
5. X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, et al., Are anchor points really indispensable in label-noise learning, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, (2019), 6838–6849.
6. X. Zhai, A. Oliver, A. Kolesnikov, L. Beyer, S4L: Self-supervised semi-supervised learning, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, (2019), 1476–1485. https://doi.org/10.1109/ICCV.2019.00156
7. D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C. A. Raffel, MixMatch: a holistic approach to semi-supervised learning, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, (2019), 5049–5059.
8. T. Miyato, S. I. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: A regularization method for supervised and semi-supervised learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (2019), 1979–1993. https://doi.org/10.1109/TPAMI.2018.2858821
9. T. Sakai, M. C. Plessis, G. Niu, M. Sugiyama, Semi-supervised classification based on classification from positive and unlabeled data, in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, (2017), 2998–3006.
10. Y. Yan, Y. Guo, Partial label learning with batch label correction, in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, **34** (2020), 6575–6582. https://doi.org/10.1609/aaai.v34i04.6132
11. N. Xu, J. Lv, X. Geng, Partial label learning via label enhancement, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, **33** (2019), 5557–5564. https://doi.org/10.1609/aaai.v33i01.33015557
12. M. L. Zhang, F. Yu, Solving the partial label learning problem: an instance-based approach, in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, (2015), 4048–4054.
13. T. Ishida, G. Niu, M. Sugiyama, Binary classification from positive-confidence data, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, Palais, (2018), 5921–5932.
14. N. Lu, G. Niu, A. K. Menon, M. Sugiyama, On the minimal supervision for training any binary classifier from only unlabeled data, preprint, arXiv:1808.10585.
15. T. Ishida, G. Niu, W. Hu, M. Sugiyama, Learning from complementary labels, in *Proceedings of the 31st International Conference on Neural Information Processing System (NeurIPS)*, Long Beach, (2017), 5644–5654.
16. X. Yu, T. Liu, M. Gong, D. Tao, Learning with biased complementary labels, in *Computer Vision—ECCV 2018*, Springer, Cham, **11205** (2018), 68–83. https://doi.org/10.1007/978-3-030-01246-5_5
17. T. Ishida, G. Niu, A. Menon, M. Sugiyama, Complementary-label learning for arbitrary losses and models, in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, **97** (2019), 2971–2980.

18. Y. T. Chou, G. Niu, H. T. Lin, M. Sugiyama, Unbiased risk estimators can mislead: A case study of learning with complementary labels, in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, **119** (2020), 1929–1938.
19. D. Liu, J. Ning, J. Wu, G. Yang, Extending ordinary-label learning losses to complementary-label learning, *IEEE Signal Process. Lett.*, **28** (2021), 852–856. https://doi.org/10.1109/LSP.2021.3073250
20. H. Ishiguro, T. Ishida, M. Sugiyama, Learning from noisy complementary labels with robust loss functions, *IEICE Trans. Inf. Syst.*, **105** (2022), 364–376. https://doi.org/10.1587/transinf.2021EDP7035
21. Y. Zhang, F. Liu, Z. Fang, B. Yuan, G. Zhang, J. Lu, Learning from a complementary-label source domain: Theory and algorithms, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2022), 7667–7681. https://doi.org/10.1109/TNNLS.2021.3086093
22. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, preprint, arXiv:1503.02531.
23. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791
24. F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.*, **65** (1958), 386–408. https://doi.org/10.1037/h0042519
25. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
26. G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
27. J. Jiang, F. Liu, W. W. Y. Ng, Q. Tang, W. Wang, Q. V. Pham, Dynamic incremental ensemble fuzzy classifier for data streams in green internet of things, *IEEE Trans. Green Commun. Networking*, **6** (2022), 1316–1329. https://doi.org/10.1109/TGCN.2022.3151716
28. L. Zhang, W. Chen, W. Wang, Z. Jin, C. Zhao, Z. Cai, et al., CBGRU: A detection method of smart contract vulnerability based on a hybrid model, *Sensors*, **22** (2022), 3577. https://doi.org/10.3390/s22093577
29. J. Jiang, F. Liu, Y. Liu, Q. Tang, B. Wang, G. Zhong, et al., A dynamic ensemble algorithm for anomaly detection in IoT imbalanced data streams, *Comput. Commun.*, **194** (2022), 250–257. https://doi.org/10.1016/j.comcom.2022.07.034

AIMS Press