硕士研究生毕业论文

Master Graduation Thesis

基于混合分布模型的噪声监督分类学习

Classification with Noisy Labels via Distribution Correction

院 系: 数学学院

专 业: 应用数学

学生姓名: 钱嘉盈

学 号: 19213286

指导教师 (职称): 任传贤 (副教授)

答辩委员会 (签名)

主席: _____

委员: _____ _____ _____

_____ _____ _____

年 月 日

论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名　：_____

日　期　：_____

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅，有权将学位论文的内容编入有关数据库进行检索，可以采用复印、缩印或其他方法保存学位论文。

学位论文作者签名:　　导师签名:

日期:　　年　　月　　日　　日期:　　年　　月　　日

摘 要

由于深度学习在各个领域的各种应用上取得的巨大成功,深度神经网络 (DNNs) 成为了计算机视觉和模式识别领域的主流算法。然而 DNNs 巨大的参数数量使得网络训练需要大量的准确标注数据。然而收集带有正确标注的大型数据集是一项十分耗时且昂贵的任务。为了客服这个困难，一种取代方式是收集来自众包平台或搜索引擎的数据。然而这类数据通常包含大量的错误标注，这种错误标注被称之为噪声标签。过去多项研究工作表明 DNNs 极易过拟合于噪声标签。为了使得 DNNs 的训练可以适应于含噪声标签的数据集，我们在图像分类任务上提出了两种能够在带有标签噪声的数据集上进行训练的方法。

第一种方法我们命名为 S^2LC（Spatial Structure mining and dynamic Label Correction)。S^2LC 分为两个组成部分，即样本选择和标签纠正。在样本选择阶段，我们基于网络的预测结果和观测标签的一致性进行选择。这一标准可以有效的过滤掉带有噪声标签的噪声样本。在标签纠正阶段，我们通过探索样本特征的空间分布，利用聚类结果纠正选择出的噪声样本的标签。经过了上述两个阶段，我们可以获得一个经过标签纠正，从而标签更为准确的数据集。我们将用经过标签纠正的数据集进行下一轮网络的训练。网络训练和标签纠正交替的进行。我们在多个虚拟和真实数据集上进行了一系列实验。实验结果证实了我们方法的有效性。

第二种方法我们命名为 DisCo(Distribution Correction)。DisCo 通过动态地赋予扩增数据集中样本权重来将带有标签噪声的数据集的噪声分布纠正为真实分布。我们提出的重要性加权方法可有有效地纠正噪声数据集的分布，同时也提出一种基于课程学习框架，通过原型学习高效的计算样本权重的方法。在数据扩增阶段，我们从纠正后的分布中通过样本插值采样虚拟样本用于扩增数据集。我们在多个虚拟数据集和真实数据集中进行了一系列的实验，实验结果证明了 DisCo 对比起当前先进方法的优越性，和 DisCo 各模块的有效性。

关键词. 深度神经网络 噪声标签 图像分类 聚类 课程学习 原型学习

Abstract

Deep learning has achieved a greate successs in many applications from severas area, and Deep Neural Networks(DNNs) has became one of the most popular algorithms in computer vision and pattern recognition. Training a DNN typically requires a mass of data with accurate annotations(clean labels) due to the large number of parameters. However, collecting a large scale dataset with clean labels is expensive and time-consuming. To overcome this problem, an alternative way is to collect data from crowd-source platforms. However, this kind of datasets inevitably contain quite a few inaccurate annotations, which are called noisy labels. Several related works have proved that DNNs are prone to overfit to the noisy labels. To make the DNNs robust to the dataset with noisy labels, we propose two noisy label learning methods for DNNs training.

The first method called S^2LC (Spatial Structure mining and dynamic Label Correction). S^2LC consists of two important phases, i.e., sample selection and label correction. For sample selection, we select samples whose network predictions are the same as the observed noisy labels as clean samples, which effectively filters out most samples with noisy labels. For label correction, we exploit the spatial structure of remaining samples and correct the labels with clustering results, based on which, more accurate pseudo labels can be obtained even under high noise level. S2LC explores an ensemble of the temporary model predictions and the spatial clustering results to supervise the training process. Extensive experimental results show that S2LC outperforms several state-of-theart methods.

The second method termed Distribution Correction(DisCo), by assigning dynamic weights to the enlarged training set, which is obtained via exploiting the data interpolation/augmentation technique. The proposed importance reweighing method is able to correct the distribution bias caused by the noisy labels, and the weights are estimated via a prototype learning phase with confident samples selection. In the data interpolation phase, virtual instances are generated by sampling from the corrected distribution. The enlarged training set is used to learn the classifier with

a weighted loss. We conducted extensive a range of experiments on both synthetic and real-world datasets with noisy labels. The results show that DisCo outperforms existing state-of-the-art deep learning methods.

Keywords. deep neural networks, noisy labels, image classification, cluster, curriculumn learning, prototype learning

目录

1 绪论

1.1 引言

近几年，深度学习与深度神经网络（DNN）在计算机视觉和模式识别领域 [1, 2, 3, 4, 5] 取得了巨大的成功。其中卷积神经网络 [6] 在许多相关任务上取得了明显的进步，比如图像分类 [2, 3, 7, 8]，物体检测 [9, 10, 1] 和实例分割 [11, 12, 13, 14] 等。深度卷积神经网络 [15] 可以充分地利用浅层，中层和深层特征。近期的研究 [16, 17] 表明卷积神经网络的深度对模型效果起到了重要的影响。在大型图像数据集 ImageNet[3] 上取得领先效果的模型均采取了层数较多的网络模型 [16, 17, 7, 18]。然而增加卷积神经网络的深度也意味着参数数目的显著增加。因此卷积网络的训练通常需要大量的有准确标注的训练数据，比如，ImageNet[3]，MS-COCO[19] 和 PASCAL VOC[20]。然而收集大量的带有准确标准的数据集是十分昂贵且花费时间的。比如，标注医学影像数据集 [21, 22, 23] 和细粒度数据集 [24, 25] 需要专家知识，因此大量的此类型数据集十分难以获得。对于收集大量训练数据的需求，一种替代的方式是收集来自众包平台 [26] 或搜索引擎 [27, 28, 29] 的数据。然而这类型的数据通常包含着大量的噪声标签，即标注错误的样本标签。由于卷积神经网络具有极强的拟合复杂函数的能力，因此在训练数据集存在噪声标签的情况下，卷积神经网络极易过拟合于噪声标签 [30, 31, 32]，Zhang 等人 [33] 通过实验说明深度神经网络可以拟合任意噪声比例的数据集。这一现象将显著降低模型的泛化性能。常用的增强模型泛化性能的方式是正则化方法，比如数据扩增 [34]，权重衰减 [35]，dropout[36] 和批正则化 [37]。然而正则化方法不能完全的解决噪声标签的问题。为了使模型的训练能够更好的适应噪声数据集，噪声标签学习成为了一个重要的研究方向。

噪声标签学习问题与离群点检测 [38, 39, 40, 41] 和异常检测 [42, 43, 44, 45, 46] 紧密相关联。在大多数情况下，标注错误的样本对于其真实所属类别，确实是以离群点或异常点的形式出现。因此许多用于处理离群点和异常点的方法也可以同样用于处理噪声标签 [47, 48]。基于离群点检测的启发，挖掘样本在特征空间上的空间结构是常用的方法 [50, 99, 100]。Wang 和 Liu[50] 使用 LOF(Local Outlier Detection) 算法 [101] 进行噪声样本检测。TopoFilter[99] 不依赖于分类器的表现，而是利用样

本在特征空间上的空间拓扑结构进行干净样本选择。NGC[100] 迭代地进行特征空间上的近邻图构建和基于近邻信息聚合的样本选择。Wang 等人 [50] 基于噪声样本通常在特征空间上以离群点的形式出现的现象，利用样本特征在特征空间的空间结构估计样本的重要性权重。Fan 等人 [49] 利用样本到每个类中心的余弦相似度计算样本的权重。然而值得注意的是，标注错误的样本并不一定是离群点或异常点。在许多现实应用场景中，易于混淆的样本通常易于被标注错误。这些噪声样本通常分布在决策边界附近，他们通常并非异常点也并非离群点（图 1(b)）。

许多相关学者提出了方法使 DNN 的训练对噪声数据集鲁棒的方法。很大一部分相关方法采用了重要性加权框架。它们通常会给每个样本赋予不同的权重以使得分类器能够更多的关注于干净样本。在最近的工作中，权重因子通过一个精心设计的权重函数进行动态的更新 [49, 50]，或者使用额外的干净数据集训练一个额外的网络进行权重因子的计算 [51, 52]。然而权重函数的设计通常缺乏理论支持，并且额外的干净数据集在现实问题中通常难以获取。重要性加权的思想也可以视作从分布的角度利用给样本赋予权重将噪声分布 $Q(X,Y)$ 纠正为潜在的真实分布 $P(X,Y)$。Liu 和 Tao[61] 在分布的角度上对样本加权系数给出了理论分析。然而这种分析是基于二分类问题的类别依赖噪声假设的。但是在现实生活中，多分类问题和实例依赖噪声是更常见的情形。

还有许多其他的噪声标签学习相关工作。Reed 等人 [53] 利用训练集观测标签和网络预测标签的凸组合进行下一轮网络的训练。Zhang 和 Sabuncu[54] 设计了一个鲁棒性的交叉熵损失函数。Wang 等人 [55] 将一个逆交叉熵项增加到经典的交叉熵损失中。Han 等人 [56] 基于样本到类原型的余弦相似度纠正样本标签。Mirzasoleiman 等人 [57] 通过近似的低秩雅可比矩阵选择干净样本子集进行模型训练。Chen 等人 [58] 使用过去的迭代的网络预测滑动平均监督下一轮网络的训练。Zheng 等人 [59] 量化了预测置信度阈值，并基于这个阈值纠正噪声样本的标签。Lee 等人 [60] 根据每一个样本到其类中心的相关度赋予不同的权重，其中类中心通过一个额外的干净数据集获得。Zhang 等人 [52] 对来自相同类的样本计算权重，并利用这些权重生成一个干净的表征。

为了对标签噪声学习问题进行理论分析和实验中数据集的构建，许多经典的噪声学习方法均采用了**类别依赖噪声假设**，即标签噪声与样本无关，仅与其真实所属

类有关。Liu 等人 [61] 在二分类问题场景中基于类别依赖噪声的假设，对于权重因子的计算进行了理论分析。Patrini 等人 [62] 通过估计噪声转移矩阵进行损失函数修正。然而，*实例依赖噪声*是现实生活中更常出现的一种噪声模式。比如在图像分类任务中，视觉上易混淆的图像更易于被错误标注。我们将二分类情形下的实例依赖噪声和类别依赖噪声可视化为图 1。我们可以观察到实例依赖噪声模式下的噪声样本通常分布在决策边界附近。而分类器通常在决策边界附近表现较差，因此实例依赖噪声是一种更为现实却也更具挑战性的场景。

基于以上讨论，从离群点检测的角度出发，我们可以通过挖掘样本的特征空间结构的方式进行标签噪声学习；从重要性加权的角度，我们需要推导出一个能够纠正噪声分布的加权方式。不管是挖掘样本的特征空间结构还是计算用于纠正噪声分布的样本权重，我们都需要一个分布模型用于拟合数据分布。混合模型是常用的拟合数据分布的假设。在很多情况下，假设样本服从某个单峰分布（比如：高斯分布）是不够准确的，无法达到很好的效果。此时采用混合模型能够更准确的对数据进行建模。因此我们可以结合混合密度估计进行噪声标签学习。同时，为了更适应现实生活中的应用场景，我们需要一个对多种噪声模式鲁棒的噪声标签学习方法。

1.2 噪声监督分类学习

考虑一个 K 分类问题。我们可以获得一个训练数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$。其中 $\mathcal{X} = \{x_i\}_{i=1}^N$ 为样本空间，$\mathcal{Y} = 0, 1, ..., K-1$ 为标签空间。训练数据集的样本独立同分布地采样自一个潜在的联合分布 P。深度学习框架下的分类问题的目标就是找到一个函数 $f(\cdot;\Theta) : \mathcal{X} \to [0,1]^K$，其中神经网络参数 Θ 应使以下经验风险最小化：

$$\mathbb{E}_{\mathcal{D}}[l(f(x;\Theta), y)] = \frac{1}{N}\sum_{i=1}^{N} l(f(x_i;\Theta), y_i) \tag{1.1}$$

其中 l 为某一损失函数。

在现实生活中，训练数据集通常会存在部分样本标注错误的情况，即所获得的训练数据集中的观测标签不一定与其真实标签一致。此时我们获得的训练数据集为 $\tilde{\mathcal{D}} = \{(x_i, y_i^o)\}_{i=1}^N$，其中的样本独立同分布地采样自噪声联合分布 Q。其中存在部分样本的观测标签与真实标签不一致，即 $y^o \neq y$，这些样本被称之为噪声样本。此时利用式 1.1进行经验风险最小化会使模型错误的拟合噪声分布，导致极差的泛化

性能。[63, 64, 65] 说明深度神经网络极易过拟合于噪声数据。

1.3 本文工作

本文将针对噪声标签学习问题，提出两个基于混合模型的方法。两个方法分别是从标签纠正和分布纠正的角度出发进行噪声标签学习。

第一种方法我们称为 S^2LC。具体地讲，我们首先在噪声数据集上预训练一个 DNN，并通过判别样本的 DNN 预测类别和样本的观测标签类别是否一致进行样本选择，针对两者标签不一致的样本，我们使用聚类结果修正这部分样本的标签，并使用经过标签纠正的数据集进行下一轮的训练。标签纠正和 DNN 的训练交替的进行。S^2LC 充分地利用了样本特征空间结构进行标签纠正，从而大幅减缓了噪声标签对 DNN 训练的影响。在多个虚拟及大规模现实数据集上的实验证明了 S^2LC 的有效性。

第二种方法我们称为 DisCo。我们从分布纠正的角度出发，提出一个结合课程学习框架通过样本加权纠将噪声数据集的噪声分布纠正为真实分布的噪声标签学习方法。我们首先推导了对样本进行重要性加权的权重系数，接下来在课程学习框架下，利用经过选择的样本训练的教师网络高效准确的计算出样本权重。同时我们还利用推广式 Mixup 对数据集进行了数据扩增。在多个虚拟及大规模现实数据集上的实验证明了 DisCo 能够超越已有的 SOTA（state-of-the-art）的方法，我们还通过一系列的消融实验证明了我们方法的样本选择，样本加权和推广式 Mixup 的有效性。

我们首先将在章节 2中介绍噪声标签研究的相关概念和相关工作，以及本文所用到的混合模型和 Mixup 算法。

在章节 3中，我们将介绍基于特征空间结构挖掘的噪声标签学习算法 S^2LC 的基本框架，算法流程和各重要组成步骤。一系列的实验结果证明，S^2LC 在人工生成的随机噪声和类别依赖噪声数据集，和大规模真实噪声数据集上均能取得优异的效果。

在章节 4中，我们将介绍基于分布纠正的噪声标签学习算法 DisCo 的基本框架，算法流程和各重要组成步骤。一系列的实验结果证明，DisCo 在更为困难的实例依赖噪声数据集上也有优越的表现。同时在两个大规模真实噪声数据集上也取得

了优异的效果。

最后，我们将在章节 5中对我们的工作进行总结并探讨未来的工作展望。

2 噪声标签学习相关概念及相关工作

在本章节，我们将首先介绍噪声标签学习工作中常考虑的几种噪声模式。接下来将介绍多种噪声标签学习的相关工作和常用算法。

2.1 噪声模式

- **随机噪声：**

  标签噪声随机的分布在数据集中，一个样本有均等的可能性被随机的标错为任意一个类。即存在一个噪声转移矩阵 $T \in [0,1]^{K\times K}$，其中 $T_{ij} = p(y^o = j|y = i)$ 是真实类别为 i 的样本被标错为 j 的概率。对于一个噪声比率 $\tau \in [0,1]$，$\forall i = j,\ T_{ij} = 1-\tau,\ \forall i \neq j,\ T_{ij} = \frac{\tau}{K-1}$。许多标签噪声学习的相关工作 [52, 66, 67, 55, 68] 会在实验中采用随机噪声虚拟数据集。

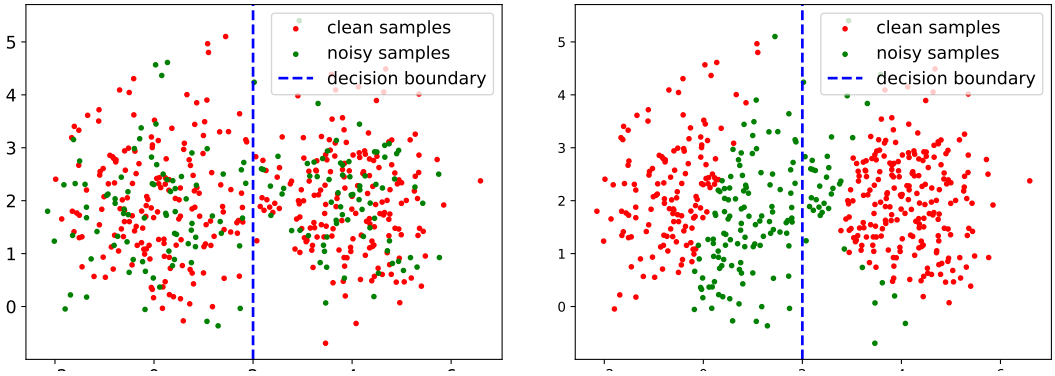
- **类别依赖噪声：**

  标签噪声仅依赖于类别，而与实例本身无关。一个样本有更倾向于被标错为特定的一个类。即原标签为 y 的样本会以概率 $p(y^o|y)$ 的概率被标错为 y^o。类别依赖噪声的噪声矩阵满足 $\forall i = j,\ T_{ij} = 1-\tau,\ \exists i \neq j, i \neq k, j \neq k,\ T_{ij} > T_{ik}$。这种噪声模式对比起随机噪声更贴近现实情况。即易于混淆的类之间易于互相标错。在 [67] 中，Tanaka 和 Ikami 等人基于类别依赖噪声利用 CIFAR10 数据集生成了虚拟噪声数据集。具体的标签噪声设定是：卡车 $\rightarrow$ 汽车，鸟 $\rightarrow$ 飞机，鹿 $\rightarrow$ 马，猫 $\leftrightarrow$ 狗。[62] 基于类别依赖噪声假设进行了理论推导，提出了一种估计噪声转移矩阵 T，并利用 T 纠正交叉熵损失函数的方法。

- **实例依赖噪声：**

  标签噪声依赖于实例。一个样本有一定的可能性被标错为实例本身易于混淆的类。即使是同一个类的实例也有可能因为实例之间的差别而被误标为不同的类。同时同一个类的不同样本也会有不同的概率被标错。实例依赖噪声可

被定义为 $\rho_{ij}(x) = p(y^o = j|y = i, x)$。在图像分类任务中，外表易于混淆的图像更易被标错。[58] 中提出一种生成实例依赖噪声的方法。图 2展示了实例依赖噪声 CIFAR10 数据集中的噪声样本。可以看出视觉上易于混淆的图像会被标错为其混淆类。实例依赖噪声样本通常分布在决策边界附近，是一种对比起随机噪声和类别依赖噪声更为实际，更难解决的噪声模式。图 1展示了虚拟数据集上的类别依赖噪声和实例依赖噪声的样本分布。可以看出实例依赖噪声样本大多分布在决策边界附近，而类别依赖噪声样本则随机的分布在两个易混淆类间。



(a) 类别依赖噪声 (b) 实例依赖噪声

图 1: 不同的噪声模式。(a) 类别依赖噪声，噪声样本在类内随机分布。(b) 实例依赖噪声，噪声样本在决策边界附近。

2.2 噪声标签对分类学习的负面影响

对于简单的问题场景和对称噪声，分类器的准确度可能会保持不受影响。对于一个二分类任务，两个类别的数据来自两个均值不同，协方差矩阵相同的高斯分布。Lachenbruch[69] 指出在这种场景下，当样本数量足够大，线性分类器的效果不受影响。事实上，在这种情况下，决策边界的改变仅与两个类的噪声率 τ_1 和 τ_2 的差 $\tau_1 - \tau_2$ 有关。这一结果在 [70] 中也有阐述。

在 [71] 中将 [69] 中考虑的场景延伸到二次判别函数分类器的情况下。对于均值不同，协方差矩阵也不相同的两个高斯分布的二分类任务。在这种情况下，即使

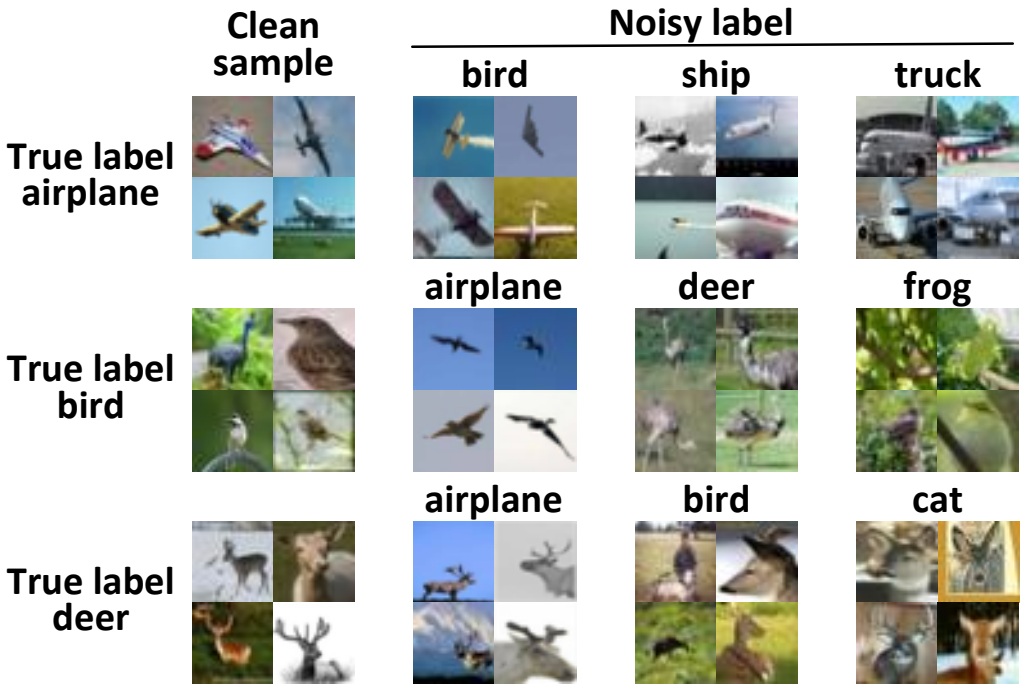


图 2: 实例依赖噪声 CIFAR10 数据集噪声样本举例

两个类别的噪声率相同（$\tau_1 = \tau_2$），分类器的表现也会受影响。分类器受影响的程度随着两个协方差矩阵的差异程度和两个噪声率的差异程度的上升而上升。Michalek 与 Tripathi[72] 和 Bi 与 Jeske[73] 提出标签噪声会影响正态判别分析和逻辑回归的参数学习。k 近邻分类器的效果同样也会受标签噪声的影响 [74, 75]。在 $k = 1$ 时，这种影响会更加明显 [76]。

对于 DNN 分类器，Zhang 和 Bengio[77] 通过实验证明 DNN 可以拟合随机标注。具体地，利用完全随机标注的训练数据集训练 DNN，神经网络可以达到 0 训练误差，然而在测试集上的预测表现却接近随机预测。实验结果证明了，DNN 具有记忆所有训练数据集的能力，即使训练集的标注是随机的。同时，作者进行实验证明了随着噪声比例的增加，DNN 的泛化能力逐渐降低。对于显式的正则化，比如权重衰减，dropout 和数据增光，实验证明它们具备一定的增强 DNN 泛化能力的效果，但依然无法充分的控制标签噪声对 DNN 泛化性能的负面影响。

2.3 噪声标签学习相关方法

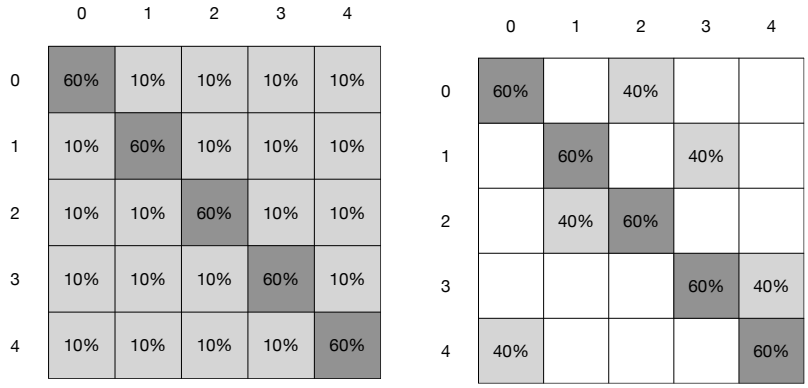
2.3.1 噪声转移矩阵

许多经典工作假设标签噪声由一个噪声转移矩阵 T 确定 [62, 78]，其中 $T_{ij} = p(y^o = j|y = i)$，如图 3。由噪声矩阵纠正的交叉熵损失定义为：

$$\begin{aligned} L &= -\frac{1}{N}\sum_{i=1}^{N}\log(\sum_{k=1}^{K}p(y=y_i|y=k)\hat{p}(y=k|x_i)) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\sum_{k=1}^{K}T_{ki}\hat{p}(y=k|x_i) \end{aligned} \tag{2.1}$$

相关工作大多聚焦于估计噪声转移矩阵 T，Patrini 等人 [62] 用一个预训练模型估计噪声转移矩阵。Hendricks 等人 [78] 用一个额外的干净数据集计算噪声转移矩阵。

Reed 等人 [53] 将噪声转移矩阵和一个正则化损失函数结合进行训练。Goldberger 等人 [79] 使用期望最大化（EM）算法寻找网络和噪声矩阵的最优参数。噪声转移矩阵在 [80, 81, 82] 等工作中得到了进一步的研究。然而现实场景中的数据集不一定符合对称噪声或非对称噪声模式，因此噪声转移矩阵方法的适用范围有限。



(a) 40% 对称噪声转移矩阵　　(b) 40% 不对称噪声转移矩阵

图 3: 噪声转移矩阵。(a) 40% 对称噪声转移矩阵。(b) 40% 不对称噪声转移矩阵。

2.3.2 鲁棒性损失函数

鲁棒性损失函数方法通常通过改变损失函数使得低预测置信度样本被较少的惩罚 [83]。Manwani 等人 [84] 证明 0-1 损失比常用的凸损失函数对噪声标签更加鲁棒。Ghosh 等人 [85] 将分类任务常用的交叉熵损失和平均绝对误差损失进行对比，并且证明平均绝对误差损失对噪声标签更加鲁棒。然而适用平均绝对误差损失进行训练通常会导致欠拟合，并且不一定适用与所有的噪声模式。Zhang 和 Sabuncu[54] 将平均绝对误差损失和交叉熵损失相结合，提出一种推广式的交叉熵损失。Wang 等人 [55] 考虑到经典的交叉熵损失的非对称性，提出了对称式的交叉熵损失，并且证明增加交叉熵损失的对称项能够有效的对抗噪声标签。Ma 等人 [86] 结合两种损失函数，利用他们能够互相促进对方的性质，提出一种鲁棒性损失函数。Ma 等人 [86] 提出现有的鲁棒性损失函数虽然具有一定的对抗噪声标签的能力，但也降低了模型的拟合能力。而 Ma 等人 [86] 提出的损失函数通过结合拟合能力强的交叉熵损失，和避免过拟合的平均绝对误差损失，达到保证模型拟合能力但又能避免过拟合与噪声标签的目的。Liu 等人 [87] 证明使用其提出的对等损失函数（peer loss function）在噪声数据集上进行训练可以得到与在干净数据集上训练得到的分类器相近的最优分类器

2.3.3 样本选择

噪声学习中的样本选择方向的目的在于获取标注正确的干净数据子集用于模型的训练。样本选择的相关工作大多基于 DNN 的*记忆效应* [88, 89, 90, 91]。在理论方面，Han 和 Yao[88] 说明在网络过拟合于噪声标签之前，神经网络的权重会保持在在初始权重的邻域内。在 [90] 和 [91] 中，网络的记忆效应在经验上得到了证实，即 DNN 趋向于先拟合简单的模式，在训练的后期过拟合于噪声样本。从 DNN 的记忆效应出发，样本选择的相关工作可以被大致分为两类：多网络协同训练和迭代式训练。

多网络协同训练的其中一个代表工作是 Decoupling[92]，Zhang 等人 [92] 提出一种同时训练两个网络的噪声标签学习方法。训练过程中仅选择两个网络预测不一致的样本对两个网络进行参数更新。另一种常用的样本选择准则是小损失准则，即将损失较小的样本视作标注正确的干净样本。MentorNet[93] 使用一个教师

网络指导学生网络进行训练。基于小损失准则，教师网络将损失较小的样本提供给学生网络进行训练。另一种使用两个网络协同训练的代表工作是 Co-teaching[94]，两个 DNN 会互相利用在另一个网络训练中损失较小的部分样本进行训练。Co-teaching+[95]j 在 Co-teaching[94] 进一步利用了 Decoupling[92] 中的不一致准则.。除此之外，JoCoR[96] 提出加入使两个网络预测一致的损失正则项来减小噪声标签的影响。然而由于需要训练额外的网络，因此大幅增加了计算复杂度。

迭代式训练不需要额外的 DNN，而是在训练过程中迭代的更新所选择的干净样本。通常所选择的样本数目会随着训练的进行而增长。ITLM[97] 在每一轮训练交替进行小损失样本选择和在选择的样本上的网络训练。INCV[98] 随机地将训练数据集划分，并采用交叉验证的方式选择损失较小的样本。

除了常用的小损失准则，许多相关工作 [50, 99, 100] 通过挖掘样本在特征空间上的拓扑结构来进行样本选择。Wang 和 Liu[50] 使用 LOF(Local Outlier Detection) 算法 [101] 进行噪声样本检测。TopoFilter[99] 不依赖于分类器的表现，而是利用样本在特征空间上的空间拓扑结构进行干净样本选择。NGC[100] 迭代地进行特征空间上的近邻图构建和基于近邻信息聚合的样本选择。但是迭代式训练通常会出现由每一轮样本选择偏差导致的累积误差。

2.3.4 样本加权

一个非常直观的解决标签噪声的方法即给每个样本赋予不同的权重。赋予较小的权重给噪声，较大的权重给干净样本，以使分类器更注重学习干净样本并降低噪声样本带来的负面影响。样本加权方法可以分为以下几种类别：重要性加权，神经网络计算，损失函数设计。

- **重要性加权：**重要性加权在领域自适应领域被提出 [102]。Liu 和 Tao[61] 将噪声训练数据视作源域，干净测试集数据视作靶域，Liu 和 Tao[61] 将重要性加权策略用于噪声标签学习领域。重要性加权的思想即从分布的角度利用给样本赋予权重将噪声分布 $Q(X,Y)$ 纠正为潜在的真实分布 $P(X,Y)$。对于噪声标签监督下的分类任务，我们的目标是学习一个函数 f，优化目标函数式

2.2：

$$\begin{aligned} R(f) &= \mathbb{E}_{(X,Y)\sim P(X,Y)}[l(f(x),Y)] \\ &= \int_x \sum_{k=1}^{K} P(X=x,Y=k)l(f(x),k)dx \\ &= \int_x \sum_{k=1}^{K} Q(X=x,Y=k)\frac{P(X=x,Y=k)}{Q(X=x,Y=k)}l(f(x),k)dx \\ &= \int_x \sum_{k=1}^{K} Q(X=x,Y=k)\frac{P(Y=k|X=x)}{Q(Y=k|X=x)}l(f(x),k)dx \\ &= \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\beta(X,Y)l(f(X),Y)] \end{aligned} \tag{2.2}$$

其中 $\beta(X,Y)=\frac{P(Y=k|X=x)}{Q(Y=k|X=x)}$ 为样本重要性加权权重。Liu 和 Tao[61] 在分布的角度上对样本加权系数给出了理论分析。然而这种分析是基于二分类问题的类别依赖噪声假设的。但是在现实生活中，多分类问题和实例依赖噪声是更常见的情形。

- **神经网络计算：** Shu 等人提出 Meta-Weight-Net[103]。Meta-weight-Net（MW 网络）可以通过数据学习到样本权重。权重函数用一个包含一个隐藏层的 MLP 进行学习。分类器参数 θ 为优化损失函数式 2.3

$$\theta^*(\phi)=\arg\min_{\theta} l^{tr}(\theta;\phi)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{V}(l_i^{tr}(\theta);\phi)l_i^{tr}(\theta), \tag{2.3}$$

其中 $\mathcal{V}(\cdot;\phi)$ 为 MW 网络，即一个包含一层隐藏层的 MLP。MW 网络用于根据分类器损失 $l_i^{tr}(\theta)$ 计算每个样本的权重 $\mathcal{V}(l_i^{tr}(\theta);\phi)$。MW 网络参数的学习是基于元学习的思想。利用少量的干净数据组成的元数据集 $\{(x_i^{(meta),y_i^{meta}})\}_{i=1}^{M}$。MW 网络参数 ϕ 的学习为优化损失函数式 2.4

$$\phi^*=\arg\min_{\phi} l^{meta}(\theta^*(\phi))=\frac{1}{M}\sum_{i=1}^{M} l_i^{meta}(\theta^*(\phi)) \tag{2.4}$$

模型训练的过程中，分类器参数 θ 和 MW 网络参数 ϕ 迭代的交替更新。然而在现实生活中，额外的干净数据集通常是难以获取的。

- **权重函数设计：** Wang 等人 [50] 基于噪声样本通常在特征空间上以离群点的形式出现的现象，利用样本特征在特征空间的空间结构估计样本的重要性权

重。Fan 等人 [49] 利用样本到每个类中心的余弦相似度计算样本的权重。然而这些方法的效果高度依赖于权重函数的设计和超参数设定。

2.3.5 标签纠正

伪标签方法 [104, 105, 53] 是一种常用的自训练方法，在噪声标签学习中也得到了广泛的应用。许多相关工作利用深度神经网络本身的泛化性能，即神经网络在训练过程中会先拟合标注正确的样本的性质 [106]，用经过适当预训练的神经网络纠正训练样本的标签 [67, 68, 53]。Reed 等人 [53] 提出用神经网络分类器预测结果和原始标签的线性组合共同的监督下一轮的网络训练。然而 Tanaka 等人 [67] 提出深度神经网络能够拟合任意数据集。因此难以控制神经网络的预训练过程是否已过拟合于噪声标签。同时利用神经网络分类器打伪标签监督下一轮的训练也有可能造成误差的累积。因此仅利用神经网络对训练样本打伪标签在噪声样本含量较高的情况下并不鲁棒。另一种标签纠正的方法是基于类原型学习 [60]。然而 CleanNet[60] 需要一个额外的经过清洗的干净数据集。而在实际问题中，一个额外的干净数据不一定能够获取。

2.4 课程学习

课程学习 [107] 将人类的学习模式应用到机器学习中。人类在学习过程中并不是随机的学习概念或知识，而是按一个有易至难的顺序进行学习。课程学习就是这种思想，根据训练样本训练的难易程度，给不同难度的样本不同的权重，一开始给简单的样本最高权重，他们有着较高的概率，接着将较难训练的样本权重调高，最后样本权重统一化了，直接在目标训练集上训练。因此课程学习可以分为两部分：一，课程设计；二，课程学习。

对于一个多分类问题，训练数据集为 $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$，其中 x_i 代表第 i 个样本，$y_i \in \mathcal{Y} = \{1, ..., K\}$ 代表 K 个类别。$L(x_i, y_i; \theta)$ 为与 m 维分类器权重参数 θ 相关的损失函数。$w \in \mathbb{R}^N$ 为 N 个样本的权重向量。课程学习的目标函数可以写作：

$$(\theta_{t+1}, w_{t+1}) = \arg \min_{\theta \in \mathbb{R}^m, w \in [0,1]^N} (\sum_{i=1}^{N} w_i L(x_i, y_i; \theta) + r(\theta) + G(w, \lambda) \tag{2.5}$$

其中 $r(\theta)$ 为权重参数 θ 的正则项。$G(w,\lambda)$ 定义了一个课程。

2.4.1 自步学习

在自步学习 [108] 中，课程定义为 $G(w,\lambda) = -\lambda\|w\|_1$，当 θ 固定时，最优 w 可写作：

$$w_i^* = \mathbf{1}(l_i \leq \lambda), \forall i \in [1, N]. \tag{2.6}$$

其中 $l_i = L(x_i, y_i; \theta)$ 为第 i 个样本的损失。$\mathbf{1}$ 为示性函数。式 2.6意味着当分类器参数 θ 固定时，损失比阈值 λ 小的样本将会被选择用于训练 $(w_i^* = 1)$，否则不会被选中 $(w_i^* = 0)$。当样本权重向量 w 固定时，分类器仅使用选中样本进行参数更新。较小的 λ 意味着只有损失较小的样本会被选择用于训练。更大的 λ，意味着更多的损失较大的样本会被选中。在自步学习中，样本的难易程度通过样本损失的大小决定。

2.4.2 MentorNet

Jiang 和 Zhou 等人 [93] 将课程学习用于噪声标签学习中，将课程设计由挑选简易样本转换为挑选干净样本，并提出了一种数据驱动的课程来取代人工设计的课程函数。MentorNet[93] 通过一个额外的经过清洗的数据集 $\mathcal{D}' = \{(x_i, y_i^o), v_i\}_{i=1}^{N'}$ 和一个额外的学生网络 $g_s(\cdot, \lambda)$ 学习课程函数 $G(w,\lambda)$ 的参数 λ。其中 y_i^o 为样本 x_i 的观测标签，$\mathbf{v}$ 为样本 (x_i, y_i) 是否干净的标注向量，即对于样本 x_i 的真实标签，如果 $y_i^o = y_i$，则 $v_i = 1$，否则 $v_i = 0$。$\mathbf{v}$ 可以视作数据集 $\mathcal{D}'$ 的最优样本权重向量。因此可以将学生网络的训练转换为训练一个二分类神经网络分类器，学生网络的输出可以视作样本干净与否的概率 w，监督信息为 $\mathbf{v}$。损失函数为交叉熵损失。即

$$\lambda^{(t+1)} = \arg\min_{\lambda}(-\sum_{i=1}^{N'} v_i \log g_s(x_i; \lambda^{(t)}) + (1 - v_i)\log(1 - g_s(x_i; \lambda^{(t)}))) \tag{2.7}$$

训练样本的权重在每一轮训练中动态更新，即固定学生网络 g_s 的参数 λ_t，更新样本权重向量 $w_i^{(t)} = g_s(x_i; \lambda^{(t)})$

2.4.3 CurriculumNet

Guo 和 Huang 等人 [109] 将课程学习用于噪声标签学习中，与 MentorNet[93] 不同的是，CurriculumNet 并未直接赋予噪声样本小权重或直接丢弃。而是设计课程按照样本难易程度将样本分成多个子集合，按照难易顺序将这些样本集合逐步加入网络训练。从而达到利用噪声样本使得模型具有更强的泛化性能，避免过拟合的目的。整个模型训练流程分为三个部分：1. 初始特征生成。2. 课程设计。3. 课程学习。

第一步首先使用标准的卷积神经网络结构，例如 Inception v2[110]，和全部的数据去学习一个初始化的模型 f。初始化模型可以将训练集图像 $x_i, i = 1, ..., N$ 投影到一个可以反映图像潜在结构和各类别图像之间的关联的深度特征空间。其中 N 为训练集图像的数目。

在课程设计阶段，CurriculumNet[109] 利用样本在特征空间上的分布来判断样本的难易程度。基于聚类假设 [111]，Guo 和 Huang 等人 [109] 提出靠近聚类中心的数据点更有可能是干净样本。利用第一步中得到的初始化模型的全连接层特征 $x_i \rightarrow f(x_i), i = 1, ..., N$，计算各类别样本特征之间的欧氏距离矩阵 $D^k \subseteq \mathbb{R}^{N_k \times N_k}, k = 1, ..., K$。

$$D_{ij}^k = \|f(x_i) - f(x_j)\|^2 \tag{2.8}$$

其中 x_i 和 x_j 均属于类别 k，K 为类别个数，$N_k, k = 1, ..., K$ 为每个类别的样本个数。D_{ij}^k 可以视作 x_i 和 x_j 之间的相似度，即越小的 D_{ij}^k 意味着 x_i 和 x_j 之间越高的相似度。

基于 $D^k, k = 1, ...K$，对类别 k 中每个图像 x_i 计算局部密度 ρ_i：

$$\rho_i = \sum_j X(D_{ij}^k - d_c) \tag{2.9}$$

其中

$$X(d) = \begin{cases} 1 & d < 0 \\ 0 & otherwise \end{cases} \tag{2.10}$$

其中 d_c 为将 $D^k \subseteq \mathbb{R}^{N_k \times N_k}$ 中的元素值从小到大排列的 $p\%$ 分位数。在 CurriculumNet 中，p 取 50 至 70 之间。ρ_i 为每一类中，到 x_i 的距离小于 d_c 的样本个

数。对于具有标注正确的干净样本，他们的外观是相似的，可以假设他们投影到特征空间上的距离也应该相近。因此干净样本的局部密度将相对较大。反之噪声样本的局部密度会较小。

对于每个图像定义距离 δ_i

$$\delta_i = \begin{cases} \min_{j:\rho_j>\rho_i}(D_{ij}^k) & if\ \exists j\ s.t. \rho_j > \rho_i \\ \max(D_{ij}^k) & otherwise \end{cases} \tag{2.11}$$

如果存在图像 x_j 使得 $\rho_j > \rho_i$，则 δ_i 为 $D_{i\hat{j}}^k$，其中 $x_{\hat{j}}$ 为最靠近 x_i 的样本。否则，如果 ρ_i 是最大的密度。δ_i 为 x_i 和离 x_i 最远的样本的特征之间的距离。可以看出，局部密度最大的样本 δ 值最大。这个样本将被选择为对应类的聚类中心。

基于上述选择的聚类中心，接下来将使用 K 均值算法将所有的数据点分类。在每一个类中，高的局部密度值意味着所有的图像特征都互相接近，即这些图像都有极高的相似性，因此这些图像大概率被正确标注。基于上述动机，CurriculumNet 使用样本的局部密度值来衡量样本为干净样本的可能性。利用样本的局部密度值，可以将样本分为从含少量噪声到含大量噪声的多个子集。

CurriculumNet 的课程学习策略的动机是将学习任务从易到难进行排序，并令模型有顺序的学习这些任务。CurriculumNet 的训练会先使用第二步中挑选出的最干净数据子集。随着训练的进行，会按顺序的逐步加入噪声数据比例较高的数据子集，最终使用完整的训练数据集进行训练。实验证明即使随着实验的进行逐步加入了噪声比例越来越高的训练样本，但是并不会降低模型的效果，反而可以作为正则化方法避免过拟合的现象。

2.5 混合模型

2.5.1 混合密度估计

考虑一个由参数 $m \in \mathbb{M} \subset \mathbb{R}^n$ 定义的分布族 $G = \{\phi(x|m)|x \in \mathbb{X}, m \in \mathbb{M}\}$。我们可以利用 K 个 $\phi(x|m)$ 组成的混合分布近似分布 $p(x) = \int \phi(x|m)$

$$p_K(x) = \sum_{k=1}^{K} \pi_k \phi(x|m_k) \tag{2.12}$$

我们可以利用 Kullback-Leibler 散度对近似误差进行度量。分布 p 和近似分布 p_k 的 Kullback-Leibler 散度定义为:

$$KL(p\|p_k) = \int p(x) \log \frac{p(x)}{p_k(x)} dx \tag{2.13}$$

Li 和 Barron[112] 证明给定一个来自某分布族的分布 p，由式 2.13定义的近似误差有界，并且与混合成分个数 k 负相关。

对于一个独立同分布采样自分布 $p(x)$ 的数据集 $\mathcal{X} = \{x_i\}_{i=1}^N$，利用 kullback-Leibler 散度衡量近似误差，寻找近似分布的目标函数为

$$\arg\min_{p_K} -\sum_{i=1}^{N} \log p_K(x_i) \tag{2.14}$$

高斯混合模型常用来拟合数据分布。假设 $\phi(x|m)$ 的形式为 $\exp(-\frac{1}{2\epsilon}(x-m)^2)$。考虑极限 $\epsilon \to 0$，则混合密度估计的目标函数转换为

$$\arg\min_{m_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \chi_{ik}(x_i - m_k)^2. \tag{2.15}$$

其中 $\chi_{ik} = 1$ 如果 $x_i \sim \phi(x|m_k)$ 否则 $\chi_{ik} = 0$。

在原型学习中，我们可以获得一个训练数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$，其中 $\mathcal{X} = \{x_i\}_{i=1}^N$ 为样本空间，$\mathcal{Y} = \in \{1, 2, ..., K\}$ 为标签空间。原型学习假设样本服从基于类别的高斯混合分布，即如果 $y_i = k$，则 $x_i \sim \phi(x|m_k)$。即式 2.15中的 χ_{ik} 已知，目标是学习类中心 $m_k, k = 1, ..., K$

2.5.2 K 均值聚类

在无监督学习任务中，我们仅能获取到无标签的数据集 $\mathcal{X} = \{x_i\}_{i=1}^N$。假设我们已知样本来自 K 个类，在原型学习的假设下，我们需要将样本分为由 K 个原型 $m_k, k = 1, ..., K$ 代表的 K 个团簇。即在目标函数式 2.15下，我们不仅需要学习类中心 m_k，还需要学每个样本点的归属 χ_{ik}。K 均值聚类的目标函数为:

$$\arg\min_{\chi_{ik}, m_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \chi_{ik}\|x_i - m_k\|^2 \tag{2.16}$$

K 均值聚类采用一种迭代的方式对式 2.16进行优化。首先初始化类中心 m_k。在第一阶段，固定 m_k，仅优化 χ_{ik}。在第二阶段，固定 χ_{ik}，仅优化 m_k。这两个阶

段交替的迭代进行，直到收敛。具体地，chi_{ik} 的更新可写做：

$$\chi_{ik} = \begin{cases} 1 & if\ k = \arg\min_j \|x_i - m_j\|^2 \\ 0 & otherwise. \end{cases} \tag{2.17}$$

m_k 的更新可写做：

$$m_k = \frac{\sum_{i=1}^{N} \chi_{ik} x_i}{\sum_{i=1}^{N} \chi_{ik}} \tag{2.18}$$

2.6 Mixup

在有监督学习中，我们的目标是学习一个函数 f，使其能够反映一个特征 x 到标签 y 的映射关系。其中 x 和 y 服从一个联合分布 $(x, y) \sim P(X, Y)$。我们需要最小化期望风险。

$$R(f) = \int l(f(x), y) dP(x, y). \tag{2.19}$$

然而我们无法得知潜在的联合分布 $P(X, Y)$，我们能够获取的只有来自 P 的一组数据集 $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$，$(x_i, y_i) \sim P(X, Y)$。利用这个训练数据集，我们可以利用一个经验分布去逼近潜在的联合分布 P。

$$P_\delta(x, y) = \frac{1}{N} \delta(x = x_i, y = y_i). \tag{2.20}$$

其中 $\delta(x = x_i, y = y_i)$ 为中心在 (x_i, y_i) 的狄拉克分布。利用式 2.20定义的经验分布，我们可以最小化经验风险。

$$R_\delta(f) = \int l(f(x), y) dP_\delta(x, y) = \frac{1}{N} \sum_{i=1}^{N} l(f(x_i, y_i)). \tag{2.21}$$

然而最小化式 2.21仅仅在 N 个样本上去逼近真实的映射关系。当我们使用一个参数数目较大的模型去学习 f（比如一个较大的神经网络），极有可能导致模型仅是记住了这 N 个样本点 [77]。从而导致模型较差的泛化性能 [113]。

为了提升模型的泛化能力，Zhang 等提出 Mixup[114] 方法用于生成虚拟样本扩增训练集。Mixup 方法利用一个通用邻近分布（generic vicinal disribution）逼近潜在的未知训练分布，即，

$$\mu(\tilde{x}, y^o | x_i, y_i) = \frac{1}{N} \sum_{j}^{N} \mathbb{E}_\lambda[\delta(\tilde{x} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j, y^o = \lambda \cdot y_i + (1 - \lambda) \cdot y_j)], \tag{2.22}$$

其中 $\lambda \sim Beta(\alpha, \alpha)$, $\alpha \in (0, \infty)$，超参数 α 控制样本对之间的插值权重。虚拟样本及其标签通过训练集的任意两个样本及其对应标签的线性组合生成：

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j, \\ y^o &= \lambda y_i + (1-\lambda) y_j, \end{aligned} \tag{2.23}$$

虚拟样本的生成方式等同于在式 2.22中定义的分布中采样获得。

然而原始的 Mixup 操作不能直接应用于噪声标签学习场景。原始的 Mixup 操作是基于训练样本进行潜在训练分布的近似，然而在训练样本存在标注错误的情况下，所得到的近似训练分布实际上是有偏差的，因此产生的虚拟样本会严重地降低模型表现。

2.7 噪声数据集介绍

为了验证噪声标签学习算法的有效性，本文将在虚拟噪声数据集和大规模真实噪声数据集上进行实验，本章将首先介绍用于生成虚拟噪声数据集的 CIFAR-10，MNIST 和 Fashion-MNSIT 数据集以及虚拟噪声的生成方式。接下来将介绍两个大规模真实噪声数据集 Clothing1M 和 FOOD101N 的概况。

2.7.1 虚拟噪声数据集

CIFAR-10 数据集分为 10 个类别。训练集中每个类别由 5000 张 32×32 的图片组成。测试集由 10000 张与训练集尺寸一致的图像组成，其中每个类别由 1000 张图像组成。Fashion-MNIST 数据集由 70,000 张来自 10 个类别的时尚产品的图像构成。其中每个类别 7,000 张图像，所有图像均为 28×28 的灰度图像。数据集分为训练集和测试集，分别包含 60,000 和 10,000 张图像。MNIST 数据集包含 70,000 张手写数字图像，其中 60,000 中用于训练，10,000 中用于测试。这些数字已经经过尺寸标准化并位于图像中心。图像是固定大小的（28×28）。CIFAR-10，MNIST 和 Fashion-MNIST 数据集都没有标注错误的噪声样本。

我们首先考虑了常用的随机噪声和类别依赖噪声设定。我们依据 [67] 中的方式在 CIFAR-10[115] 和 Fashion-MNIST[116] 生成随机噪声和类别依赖噪声虚拟数据集。

对于 CIFAR-10 和 Fashion-MNIST 随机噪声数据集和给定的噪声比例 p，我们以 p 的概率给样本赋予一个随机的标签，以 $1-p$ 的概率保留前真实标签。对于类别依赖 CIFAR-10 数据集，我们以 p 的概率将类别为卡车的样本标注为汽车，类别为鸟的样本标注为飞机，类别为鹿的样本标注为马，类别为猫的样本与类别为狗的样本互换标签。对于类别依赖 Fashion-MNIST 数据集，我们以 p 的概率将类别为靴子的样本标注为运动鞋，类别为运动鞋的样本标注为便鞋，将类别为套头衫的样本标注为衬衫，将类别为外套和连衣裙的样本互换标签。

Chen 和 Ye[58] 基于更"难"的样本更易于被错误标注的直觉 [117, 118]，提出了一种在标注正确的干净数据集上生成虚拟实例依赖噪声的算法。对于一个干净的数据集 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{N}$，Chen 和 Ye 首先在 $\mathcal{D}$ 上训练一个 DNN 分类器 T 次，得到 T 组对数据集中样本的预测概率向量。如果对某个样本，网络以较高的概率将其预测为非其所属类的类别，则证明这个样本是较难分类的样本。每个样本的错误标注得分 N_i 和其潜在的噪声标签 $\tilde{y}_i$ 可以计算为：

$$\begin{aligned} S &= \sum_{t=1}^{T} S^t/T \in \mathbb{R}^{N\times K} \\ N_i = \max_{k\neq y_i} S_{i,k},\ \tilde{y}_i &= \arg\max_{k\neq y_i} S_{i,k} \end{aligned} \tag{2.24}$$

其中 $S^t=\hat{p}(y|x_i)\in\mathbb{R}^K,\ i=1,...,N$ 为 DNN 在第 t 次迭代中对每个样本的预测概率向量。若要生成噪声比例为 $p\%$ 的噪声数据集，则将每个样本的错误标注得分 N_i 由高至低排序的前 $p\%$ 的样本的标签替换为其对应的潜在噪声标签 $\tilde{y}$。实例依赖噪声生成算法总结为算法 1。

图 2展示了噪声 CIFAR-10 数据集中的噪声样本与其对应的真实标签和噪声标签。可以看出噪声样本在视觉上均易与其噪声标签混淆。

2.7.2 大规模真实噪声数据集

Clothing1M[30] 包含了一百万张来自多个线上购物网站的服装图像。Xiao 和 Xia[30] 定义了 14 个类，分别为 T 恤，夹克，羽绒服，套装，披肩，连衣裙，背心和内衣。对收集的图像的标注方式为如果图像的描述文字仅包括上述 14 个类中的某一个关键字，则标注为对应类，否则丢弃这张图像以避免歧义。数据集被划分为训练集，验证集和测试集，分别包含 47,570，14,313 和 10,000 张图像。剩余的图像

Algorithm 1 实例依赖噪声生成

输入: 干净样本数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$，生成噪声数据集噪声比例 p，迭代次数 T。

输出: 实例依赖噪声数据集 $\bar{\mathcal{D}} = \{(x_i, y_i^o)\}_{i=1}^N$.

初始化 DNN。
for $t = 1$ to T **do**
  for mini-batch $\{(x_i, y_i)\}_{i \in \mathcal{B}}$ **do**
    使用交叉熵损失在 $\{(x_i, y_i)\}_{i \in \mathcal{B}}$ 上训练 DNN。
  end for
  记录 DNN 输出预测概率向量 $S^t = \{\hat{p}(y|x_i)\}_{i=1}^N \in \mathbb{R}^{N \times K}$
end for
用式 2.24计算 N_i 和 $\tilde{y}_i$。
计算指标集 $\mathcal{I} = \{p\% \arg\max_{1 \le i \le N} N_i\}$。
替换标签 $y_i^o = \tilde{y}_i \ if \ i \in \mathcal{I} \ else \ y_i^o = y_i$。

经过人工标注，组成一个干净数据集。在我们的实验中不需要用到干净数据集。此数据集的类别样本数和噪音分布都及其不平衡。噪声样本主要集中在部分类别，且不同类别的噪声比例也不一样。Clothing1M 的标签准确度估计为 61.54%。

Food-101N[60] 包含 310,000 张来自 Google，Bing，Yelp 和 TripAdvisor 的图像，图像的类别标注方法和 Food-101[119] 一致。Food-101N 没有收集来自 foodspotting.com 的图像，即避免了出现和 Food-101 重复的图像。Food-101N 的标注准确度估计为 80%。Lee 和 He[60] 人工标注了 55,000 张训练集图像和 5,000 张图像用于检验噪声样本检测效果。实验中测试集采用 Food-101 的测试集。

2.8 本章总结

本章中，我们首先介绍了噪声标签学习中常用的噪声模式假设和噪声标签对分类学习的负面影响。接下来我们分别从噪声转移矩阵，鲁棒性损失函数，样本选择，样本加权，标签纠正五个角度对噪声标签学习相关工作进行了总结。接下来我们介绍了分类学习中常用的四个算法与模型，即课程学习，Mixup 和混合密度估计与 K 均值聚类。最后我们介绍了常用的两类噪声标签数据集，虚拟噪声数据集和大规模

真实噪声数据集。

3 基于特征空间结构挖掘的噪声标签学习算法

3.1 引言

为了适应实际应用场景中的机器学习问题，更好地利用带有标签噪声的数据集，许多对抗标签噪声的机器学习算法被提出。其中样本选择和标签纠正是两大常见的研究方向。然而通过我们在章节 4.2.1和章节 2.3.5中的介绍，前人的样本选择方法大多存在对超参数敏感和计算复杂度过大的问题。同样的，前人的标签纠正方法也存在对高噪声情形不鲁棒的缺点。并且部分标签纠正方法需要一个额外的经过清洗的数据集，而这在实际应用中通常难以获取。

本章中，我们提出了一种基于样本选择和标签纠正的噪声学习算法 S^2LC。我们的算法可以有效地解决上述前人工作在样本选择和标签纠正中存在的缺点。具体地，在样本选择方面，我们将经过适当预训练的 DNN 分类器预测结果与样本观测标签（与真实标签有可能不一致）一致的样本视作干净样本，反之视作噪声样本。我们的样本选择策略是基于 DNN 的记忆效应 [106]，即 DNN 的训练具有先拟合干净样本，再拟合噪声样本的特性。因为训练数据集存在一部分的干净样本，如果我们在预训练 DNN 时采用早停策略和较大的学习率 [67]，DNN 分类器将在干净样本上表现良好，且尚未过拟合于噪声样本。因为我们无法获取训练样本的真实标签，因此我们将 DNN 的预测结果视作训练样本标签的一个近似，基于经过预训练的 DNN 分类器在干净样本上表现良好的现象，我们将 DNN 分类器预测结果与观测标签一致的样本视作干净样本。我们的样本选择策略不需要额外的超参数和额外的网络训练。

在样本选择之后，我们使用聚类结果纠正剩余样本的标签。我们的样本纠正方法是受到 [111] 中提出的聚类假设和离群点检测工作的启发。在噪声数据集上经过适当预训练的 DNN 提取的特征会呈现基于类别的团簇状分布，其中标注错误的噪声样本会以离群点的形式出现，并在特征空间中将靠近其真实所属类别。图 4展示了这种现象。我们可以观察到同一类样本的特征呈团簇状分布。但是每一个团簇都包含一部分的来自另一个类的样本（展示为不同的颜色）。这些样本有很大的可能

性为标注错误的噪声样本，且实际上属于其在空间上靠近的团簇对应的类别。如图4的右侧所展示，部分被标注为类别 1 的样本实际属于类别 9。

混合模型是一种常用于拟合数据分布的模型，基于高斯混合模型的聚类方法，比如 K 均值聚类和高斯混合聚类，可以充分地利用空间结构进行聚类。利用聚类方法，图 4右侧所展示的噪声样本可以通过聚类结果将标签纠正为其真实所属类。上述现象启发我们使用聚类方法进行标签纠正。

我们利用 DNN 的记忆效应和样本特征空间分布，提出了一种噪声标签学习算法 S²LC（**S**patial **S**tructure mining and dynamic **L**abel **C**orrection）。S²LC 通过 DNN 分类器的预测结果与观测标签的一致性进行样本选择并利用聚类结果纠正噪声标签。在样本选择和标签纠正两个步骤后，我们使用纠正后的标签和 DNN 的预测标签共同的监督下一轮的训练。

S²LC 有以下几点优势：1）不需要对于数据集噪声的任何先验知识（比如，噪声比例等）。2）不需要训练额外的 DNN。3）样本选择中不引入额外的超参数。

我们的工作贡献可以总结为以下几点：

- 我们提出一种噪声标签学习算法 S²LC。S²LC 包含两个阶段，即样本选择和标签纠正。样本选择的准则为 DNN 分类器的预测与观测标签的一致性。样本选择阶段之后，我们利用聚类结果纠正剩余的样本标签。以上两个阶段和 DNN 分类器的预测交替进行。

- 为了提高基于样本在特征空间上的团簇结构的聚类准确性，我们加入 $(C+1)$ 元组损失 [120] 进行判别性特征学习。为了避免噪声标签带来的负面影响，我们仅在选择出来的干净样本上优化 $(C+1)$ 元组损失。

- 我们在真实和虚拟噪声数据集上进行了一系列实验。实验结果证明了 S²LC 各阶段的有效性，并且效果超越了现有的噪声标签学习深度方法。

3.2 S²LC 算法描述

考虑一个 K 分类问题，我们的目标是训练一个 DNN 分类器 $F(\cdot;\theta)$ 来预测每一个样本 x_i 的标签 y_i，其中 θ 为 DNN 的参数。对于分类问题，我们通常优化交

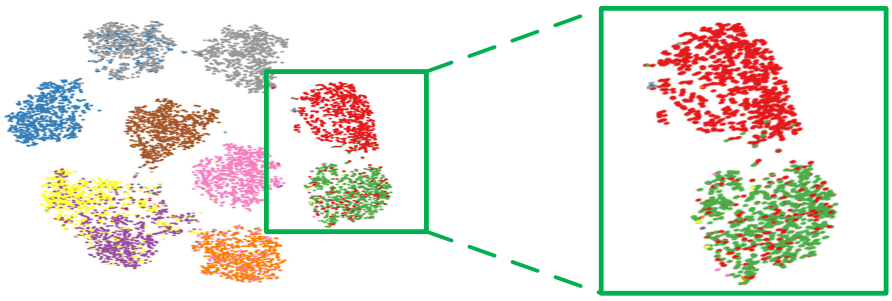


图 4: 在包含 20% 的类别依赖噪声的 CIFAR10 数据集上利用交叉熵损失函数训练 30 轮迭代的特征分布（利用 t-SNE 图进行可视化）。部分类别 9 的样本被错误标注成类别 1。左图展示的是所有训练样本的特征分布。不同的颜色代表样本不同的观测标签。右图我们展示了观测标签为类别 1（标注为红色）和类别 9（标注为绿色）的样本特征分布。在绿色团簇中出现的红色样本点即为真实标签为 9 但是被错误标注为 1 的噪声样本。

叉熵损失。

$$\theta^* = \arg\min_{\theta} \sum_{1}^{N} l_{ce}(F(x_i;\theta), y_i) \tag{3.1}$$

其中 l_{ce} 代表交叉熵损失函数。在干净数据集下通过优化式 3.1进行训练能达到很好的效果。然而在训练集中存在标签噪声的情况下，即部分样本的观测标签 y_i^o 和其真实标签 y_i 不一致。此时直接优化式 3.1会极大的降低模型的泛化性能。

S²LC 的模型框架如图 5所示。S²LC 包括两个步骤：1）DNN 训练。2) 样本选择和标签纠正。这两个步骤交替地进行。我们在每一轮网络训练迭代之后进行样本选择并利用聚类结果纠正剩余样本的标签。为了进一步地得到更为准确的聚类结果，我们在选择出的样本上优化 $(C+1)$ 元组损失。这一操作使得样本特征具有判别性，即同类样本特征靠近，异类样本特征远离。纠正后的标签和当前迭代的 DNN 预测标签将用于监督下一轮的 DNN 训练。

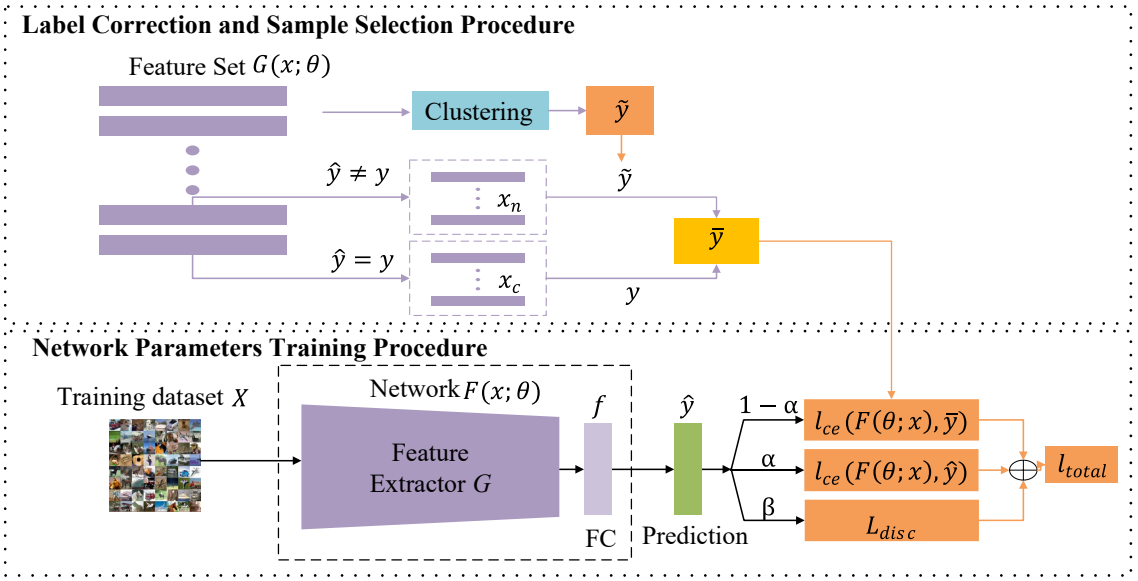


图 5: S²LC 的方法示意图。DNN 训练和样本选择与标签训练交替进行。

3.2.1 干净样本选择

我们基于 DNN 的预测结果和样本观测标签的一致性将训练数据集分为干净样本集 X_c 和噪声样本集 X_n。

$$\begin{aligned} X_c &= \{x_i | F(x_i;\theta) = y_i^o,\ i = 1, ..., N\} \\ X_n &= \{x_i | F(x_i;\theta) \neq y_i^o,\ i = 1, ..., N\} \end{aligned} \tag{3.2}$$

我们首先使用交叉熵损失和训练集观测标签预训练 DNN 分类器。依照 [67]，我们在预训练过程中使用了较高的学习率和早停策略以避免过拟合问题。在 S²LC 中，样本选择和 DNN 训练是交替进行的，因此这两个步骤可以互相促进彼此。

3.2.2 标签纠正

在样本选择之后，我们以干净样本各类别的样本特征的均值作为类中心，初始化 K 均值聚类，并对所有的样本特征进行聚类。我们利用聚类结果纠正剩余样本的标签。为了提高我们的方法的效率，在每一轮 DNN 训练之后，我们会首先使用主成分分析法对 DNN 深度特征进行降维。接下来，我们使用 K 均值聚类方法对训练集所有样本进行聚类。我们使用匈牙利算法 [121] 将聚类结果和 DNN 分类结果进行标签对齐。另一种标签对齐的做法是将聚类结果和样本观测标签进行标签

对齐。然而基于 DNN 的记忆效应，经过恰当预训练的 DNN 在选择出的噪声样本 X_n 上的预测结果比观测标签更加可信。经过标签对齐后，我们用得到的聚类标签纠正 X_n 中样本的标签。对于选择出的干净样本 X_c，观测标签会较为可信。我们将 DNN 分类器 $F(x_i;\theta)$ 的预测结果记为 $\hat{y}_i$，聚类结果记为 $\tilde{y}_i$。在每一轮 DNN 训练之后，我们对训练数据集的标签进行更新。

$$\bar{y}_i = \begin{cases} y_i^o,\ if\ x_i \in X_c \\ \tilde{y}_i,\ if\ x_i \in X_n \end{cases} \tag{3.3}$$

3.2.3 判别性特征学习

基于混合模型的聚类准确度高度依赖于样本在特征空间上的判别度，即同类特征互相靠近，异类特征互相远离。我们利用 $(K+1)$ 元组损失 [120] 提高特征的判别度。定义 $(K+1)$ 元组为 $\{x_i, x_i^+, x_i^1, ..., x_I^{K-1}\}$，其中 x_i^+ 和 x_i 有相同的观测标签，$\{x_i^k\}_{k=1}^{K-1}$ 有不同的观测标签。$(K+1)$ 元组损失 L_{disc} 定义为

$$L_{disc} = -\sum_{i\in\{i|x_i\in X_c\}} \log \frac{\exp(f_i^T f_i^+)}{\exp(f_i^T f_i^+) + \sum_{k=1}^{K-1}\exp(f_i^T f_i^k)} \tag{3.4}$$

其中 $f_i \in \mathbf{R}^d$ 为 d 维特征向量。f_i^+ 和 f_i^k 为样本 x_i^+ 和 x_i^k 对应的特征向量。

Zhang 和 Yao 指出 [92] 噪声标签对表征学习的影响比对分类器学习要大。因此在噪声数据集上直接优化 $(K+1)$ 元组损失会极大地降低特征表示的质量。从而导致聚类和分类的效果降低。然而，因为 X_c 中的噪声比例要远低于整个训练集的噪声比例，因此我们仅在 X_c 上优化 $(K+1)$ 元组损失以避免噪声样本带来的负面影响。

3.2.4 优化目标函数

由于聚类方法是一种无监督学习的方法，因此依然有可能错误分类部分样本。另一方面，训练数据集中依然存在部分干净样本，并且 DNN 分类器是经过有监督训练的。因此我们不能简单的丢弃 DNN 分类器对样本的预测结果。在 S²LC 中，DNN 分类器的预测结果也被利用于监督分类器的下一轮训练。具体地，在预训练

过程之后，优化目标函数为：

$$L_{cls} = (1-\alpha)\sum_{i=1}^{N} l_c e(F(x_i;\theta),\bar{y}_i) + \alpha\sum_{i=1}^{N} l_c e(F(x_i;\theta,\hat{y}_i{}^{(t-1)})) \tag{3.5}$$

其中 $\hat{y}_i{}^{(t)}$ 代表第 t 轮迭代的 DNN 分类器预测结果。α 控制 DNN 预测结果在训练中的重要性。

总体的优化目标函数为：

$$L_{total} = L_{cls} + \beta L_{disc} \tag{3.6}$$

其中 β 控制判别性特征学习的重要性。

S^{2}LC 可以总结为算法 2。S^{2}LC 包含三个阶段，即预训练阶段，判别性特征学习阶段和标签纠正阶段。在预训练阶段，我们使用训练数据集和样本观测标签预训练一个 DNN 分类器。在判别性特征学习阶段，我们开始将 $(C+1)$ 元组损失（式 3.4）加入优化目标函数。在标签纠正阶段，我们使用纠正后的标签优化总体目标函数（式 3.6）。

3.3 实验

我们在虚拟数据集和大规模真实噪声数据集上验证 S^{2}LC 的效果。我们将介绍我们对比的相关方法和实验设定。接下来展示实验结果和相关分析。

3.3.1 对比方法

我们将 S^{2}LC 与以下代表方法进行对比。分别为，交叉熵损失（CE），Forward[62]，MAE[85]，GCE[54]，Joint Optimiazation（Joint Optim.）[67] 和 Taylor-CE[122]。除了 GCE 和 Joint Optim.，所有方法的超参数选择都基于验证集的表现。GCE 和 Joint Optim. 的超参数设定与其原论文一致。在消融实验中，我们汇报了加入和不加入判别性特征学习的 S^{2}LC 的测试准确度。我们将加入判别性特征学习的 S^{2}LC 记为 S^{2}LC + D。

3.3.2 实验细节

在 CIFAR-10 上生成的随机噪声数据集和类别依赖噪声数据集上，我们参照 [67] 中的设定，使用 PreAct ResNet-32[123] 网络。我们进行均值相减作为数据预处

Algorithm 2 S^2LC

输入: 训练集 $\mathcal{D} = \{(x_i, y_i^o)\}_{i=1}^N$,

总迭代次数 T, 判别性特征学习阶段 T_d, 标签纠正阶段 T_c, 学习率 ϵ.

参数: 网络参数 θ。

输出: $F(x;\theta)$.

初始化网络 $F(x;\theta)$。
for $t = 1$ to T **do**
  if $t < T_d$ **then**
    从训练集中采样采样 (X, Y)。
    更新 $\theta^{(t+1)} = \theta^{(t)} - \epsilon\nabla l_{ce}(F(x;\theta^{(t)})), Y)$。
  else
    if $t < T_c$ **then**
      将训练集样本 $\mathcal{X}$ 分为干净样本集 $\mathcal{X}_c$ 和噪声样本集 $\mathcal{X}_n$。
      更新 $\theta^{(t+1)} = \theta^{(t)} - \epsilon\nabla L_{total}(F(x;\theta^{(t)}), y)$。
    else
      使用 PCA 进行特征降维。
      在当前迭代进行 K 均值聚类。
      利用式 3.3纠正 $\mathcal{X}_n$ 中样本的标签。
      更新 $\theta^{(t+1)} = \theta^{(t)} - \epsilon\nabla L_{total}(F(x;\theta^{(t)}), \bar{y})$。
    end if
  end if
end for

理。数据扩增操作为水平随机翻转和在每一边补 4 个像素点后的 32×32 随机切割。在网络优化参数设定方面，我们使用动量设置为 0.9 的 SGD 优化器，权重衰减设置为 10^{-4}，批大小设置为 128。在 Fashion-MNIST 上生成的随机噪声数据集和类别依赖噪声数据集上，我们使用 ResNet-18[2] 网络，其他的实验设定与 CIFAR10 一致。

对于 Clothing1M 数据集，我们参照 [67] 中的设定，使用在 ImageNet 上预训练的 ResNet-50[2] 作为核心网络。数据预处理操作为，先调整图像大小为 256×256，接下来进行均值相减，最后切割中间的 224×224 像素的部分。我们使用动量设置为 0.9 的 SGD 优化器，权重衰减设置为 10^{-3}，批大小设置为 32。

3.3.3 对比实验结果与分析

CIFAR10：表 1展示了在类别依赖噪声 CIFAR-10 数据集上的实验结果。S^2LC 和 S^2LC + D 取得了最优的测试集准确度。其中 Joint Optim.[67] 在噪声比例小于 40% 的情况下能够取得和我们的方法相近的效果。然而，当噪声比例从 30% 增加到 40% 的时候，Joint Optim.[67] 的测试集准确度下降了 3.06%，而我们的方法仅下降了 1.08%。S^2LC 与 Joint Optim.[67] 均采用了标签纠正的思想，我们的方法与 Joint Optim.[67] 的不同点在于我们不仅仅依赖于 DNN 分类器的预测结果，还额外利用了聚类的结果。实验结果证明了我们的标签纠正方法在高噪声比例情形下比 Joint Optim.[67] 更为鲁棒。同时我们的方法也超过了两个鲁棒性损失函数方法：GCE[54] 和 Taylor-CE[122]。且随着噪声比例的升高，我们的方法所获得的提升越明显。这证明了我们方法对比起更改损失函数，在噪声标签学习上更有效。

使用噪声数据集训练 DNN 模型通常易于过拟合于噪声标签，导致较差的泛化性能。具体表现为模型训练最后一轮迭代的测试准确度远低于模型验证集效果最优的一轮迭代的测试集准确度。基于以上现象，我们将最优验证集准确度模型的测试集准确度记为 BEST，最后一轮迭代的测试集准确度记为 LAST。如果这两个结果相差不大，则证明模型具有避免过拟合的效果。表 3展示了我们的方法和交叉熵训练的最优验证效果模型测试准确度和最后一轮迭代模型准确度。我们可以看出，仅使用交叉熵在噪声数据集上训练时，最后一轮迭代的测试准确度会显著下降，并且随着噪声比例的增加，这种下降将会愈加明显。然而 S^2LC 和 S^2LC + D 不仅在两

个结果上都超出了 CE，并且最后一轮迭代的测试准确度并未显著下降。这证明了我们的模型不仅可以提升在噪声标签上的训练效果，并且具有避免过拟合的能力。

表 2展示了在随机噪声 CIFAR10 数据集上的实验结果。实验结果再次证明了 S^2LC 在各噪声比例下均能取得最优越的方法。我们的方法具有在不同噪声模式和不同噪声比例下的鲁棒性。

方法	类别依赖噪声			
噪声比例 (%)	0.1	0.2	0.3	0.4
CE	90.66	89.85	88.5	86.29
GCE	89.28	84.63	87.23	80.5
Taylor-CE	87.34	85.02	79.37	72.65
Joint Optim.	90.74	90.10	89.43	86.37
S^2LC	<u>91.50</u>	<u>90.91</u>	<u>89.73</u>	<u>88.65</u>
S^2LC + D	**91.84**	**91.24**	**90.58**	**89.35**

表 1: 类别依赖噪声 CIFAR-10 数据集实验结果，最优结果为**粗体**显示，次优结果为<u>下划线</u> 展示。

FASHION-MNIST：表 5和表 4展示了在类别依赖噪声和随机噪声 FASHION-MNIST 数据集上的实验结果。从实验结果可以看出，在两个噪声模式和所有的噪声比例下我们的方法均取得了最优的效果。值得注意的是，即使在更高的噪声比例下，各方法在随机噪声下的效果也普遍高于类别依赖噪声。这证明了类别依赖噪声是一个更难的噪声标签学习场景。而我们的方法在类别依赖噪声情景也取得了明显的提升，证明了我们的方法即使在较为困难的类比依赖噪声情景下也具有良好的鲁棒性。在类别依赖噪声情形下，除了我们的方法以外，专门针对类别依赖噪声设计的方法 Forward[62] 取得了最优的效果。而 Forward 在随机噪声情形下却并未取得优异的效果（在 0.8 的噪声比例下，测试准确度比 Truncated 低 0.74%）。这证明了 Forward 虽然对类别依赖噪声效果较好，但在不同的噪声模式下不具有稳定的

方法	随机噪声				
噪声比例 (%)	0.1	0.2	0.3	0.4	0.5
CE	85.98	84.06	83.49	81.88	79.31
GCE	89.07	88.09	87.8	86.5	84.91
Joint Optim.	90.18	88.75	88.36	85.18	<u>86.52</u>
S^2LC	**91.08**	<u>90.00</u>	<u>89.35</u>	<u>87.86</u>	86.49
$S^2LC + D$	<u>90.95</u>	**90.58**	**89.76**	**88.72**	**87.35**

表 2: 随机噪声 CIFAR-10 数据集实验结果。

method		类别依赖噪声			
noise rate(%)		0.1	0.2	0.3	0.4
CE	BEST	90.66	89.85	88.5	86.29
	LAST	88.9	87.57	84.5	81.84
S^2LC	BEST	<u>91.50</u>	<u>90.91</u>	<u>89.73</u>	<u>88.65</u>
	LAST	91.14	90.74	89.4	88.11
$S^2LC + D$	BEST	**91.84**	**91.24**	**90.58**	**89.35**
	LAST	91.61	91.07	90.2	88.82

表 3: 类别依赖噪声 CIFAR10 数据集最优验证效果模型的测试准确度（记为 BEST）和最后一轮迭代模型的测试准确度（记为 LAST）。

表现。而我们的方法在两个噪声模式下均超过了 Forward，证明了我们的方法对不同噪声模式的鲁棒性。

方法	类别依赖噪声			
噪声比例 (%)	0.1	0.2	0.3	0.4
CE	94.06 ± 0.05	93.72 ±0.14	92.72±0.21	89.82±0.31
MAE	74.03±6.32	63.03±3.91	58.14±0.14	56.04±3.76
Forward	94.33±0.10	94.03±0.11	93.91±0.14	93.65±0.11
GCE	93.53±0.17	93.36±0.07	92.76±0.14	91.62±0.34
Taylor-CE	90.25±0.09	90.31±0.13	88.31± 0.15	87.38±0.25
S^2LC	94.99±0.32	94.52±0.05	94.39±0.08	94.27±0.13
S^2LC+D	**95.09±0.09**	**94.91±0.06**	**94.75±0.05**	**94.45±0.16**

表 4: 类别依赖噪声 FASHION-MNIST 数据集实验结果。

Clothing1M：表 6展示了在大规模真实数据集 Clothing1M 上的实验结果。我们方法超越了其他的对比方法，S^2LC + D 和 S^2LC 分别取得了最优和次优的结果。Forward[62] 利用了额外的包含 50,000 张图像的标注正确的干净数据集进行噪声转移矩阵的估计，而额外的干净数据集通常在现实中难以获取。我们的方法不需要额外的干净数据集，对比起 Forward 依然取得了 2.76% 的提升。

3.3.4 消融实验

我们将首先通过实验结果分析 S^2LC 的三个重要组成部分，即标签纠正，样本选择和判别性特征学习的有效性。最后将分析 S^2LC 的超参数的鲁棒性。

标签纠正：为了展现 S^2LC 中标签纠正的效果。我们在 30% 随机噪声 CIFAR10 数据集上进行实验，并将训练过程中 DNN 分类器预测准确度和 K 均值聚类准确度的变化展示在图 6(a) 。我们可以观察到，在训练的初期，K 均值聚类的效果优于

方法	随机噪声			
噪声比例 (%)	0.2	0.4	0.6	0.8
CE	93.24±0.12	92.09±0.18	90.29±0.35	86.20±0.68
MAE	80.39±4.68	79.30±6.20	82.41±5.29	74.73±5.26
Forward	93.64±0.12	92.69±0.20	91.16±0.16	87.59±0.35
Truncated	93.21±0.05	92.60±0.17	91.56±0.16	88.33±0.38
Taylor-CE	89.96 ±0.11	88.97 ± 0.28	87.07± 0.31	78.95±0.47
S^2LC	93.94±0.21	93.27 ±0.13	91.79±0.01	87.9±1.75
S^2LC+D	**94.48±0.09**	**93.58±0.36**	**92.11±0.15**	**88.58±0.60**

表 5: 随机噪声 FASHION-MNIST 数据集实验结果。

DNN 分类器，随着训练的进行，DNN 分类器的效果逐渐上升。这个现象指出了我们同时利用 K 均值聚类结果和 DNN 分类器预测结果的合理性。图 6(b)展示了纠正后标签的整体标签准确度随着训练过程的变化。我们可以看出纠正后的标签准确度稳定的上升，并且显著的高于原噪声数据集的观测标签准确度。这个结果指出纠正后的标签能够比原噪声标签更好的监督模型的训练。

样本选择：为了衡量样本选择的效果，我们使用了标签精确度（LP）和标签召回度（LR）两个指标 [94]，定义为式 3.7。

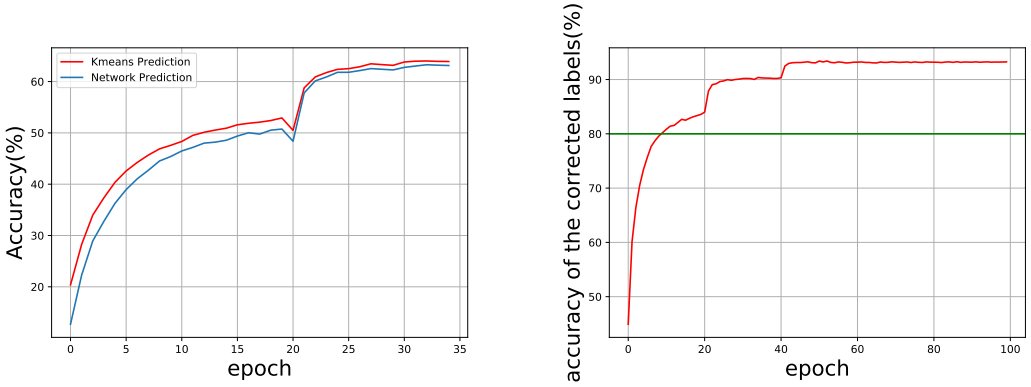
$$\begin{aligned} LR &\triangleq \frac{|\{(x,y) \in X_c : y^o = y\}|}{|\{(x,y) \in X_c : y^o = y\}|}, \\ LP &\triangleq \frac{|\{(x,y) \in X_c : y^o = y\}|}{|X_c|}, \end{aligned} \tag{3.7}$$

从式 3.7中可以看出，LP 代表 X_c 中干净样本的比例，LR 代表 X_c 中干净样本占训练集中所有干净样本的比例。

从图 7(a) 和图 7(b)中可以看出，随着训练的进行，LP 和 LR 逐步上升。其中 LR 在第 20 次迭代之后持续地超过 0.9，且 LP 的最小值也大于 0.96。这证明了我

方法	测试准确度 (%)
CE	68.94
Forward	69.84
Joint Optim.	72.16
S^2LC	72.41
$S^2LC + D$	**72.96**

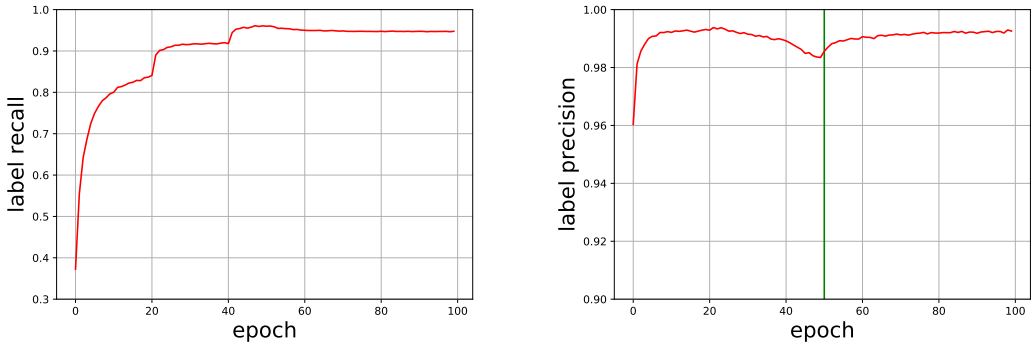
表 6: Clothing1M 实验结果



(a) (b)

图 6: 标签纠正结果可视化。(a) DNN 分类器预测结果和 K 均值聚类预测结果准确度对比。(b) 纠正后标签的准确度，水平绿线代表数据集噪声率。

们的样本选择策略能够精确且全面地选择出干净样本。我们还观察到在第 20 次迭代时 LP 达到峰值但随后下降，这个现象可能是由过拟合导致。但是在第 50 次迭代之后，LP 再次开始平稳增长，这是由于在第 50 次迭代时我们开始了标签纠正阶段的训练，即交替地进行网络更新和标签纠正。这一现象证明了 S^2LC 交替训练的有效性。



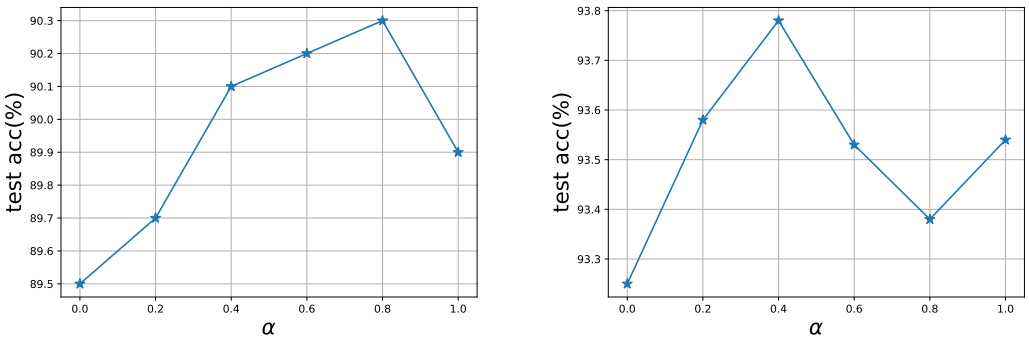
(a) (b)

图 7: 样本选择结果可视化。(a) 标签召回度。(b) 标签精确度。(b) 中的垂直绿线代表第 50 次迭代。第 50 次迭代之后我们开始使用纠正后的标签进行训练。此时标签精确度开始再次提升。

判别性特征学习：在 CIFAR10，FASHION-MNIST 和 Clothing1M 的实验结果中可以看出，S^2LC + D 对比起 S^2LC 在不同数据集，不同噪声模式和不同噪声比例下均能取得更为优越的效果。从表 2中可以看出来，在高噪声比例下，判别性特征学习所带来的提升会更加明显。在 10% 的噪声比例下，加入判别性特征学习并未带来明显的提升，但是在 50% 的噪声比例下，判别性特征学习带来了 0.86% 的提升。出现这种现象的原因可能是，当噪声比例较小时，DNN 仍具有提取具有判别性的特征的能力。此时 $(K+1)$ 元组损失的效果并不明显。而当噪声比例较高时，DNN 表征学习的质量将严重受噪声标签的影响，因此无法提取高质量的具有判别性的特征。此时 $(K+1)$ 元组损失将促进 DNN 的表征学习效果，使 DNN 能够提取具有判别性的特征。

超参数敏感性分析：权重系数 α 决定了在 S^2LC 训练的标签纠正阶段，DNN 预测结果在模型训练中的重要程度。当 $\alpha = 0$，模型的训练仅由聚类结果监督。当 $\alpha = 1$，

聚类结果被完全抛弃。图 8记录了不同的 α 取值在噪声 FASHION-MNIST 数据集上的测试集准确度。我们可以看出，$\alpha = 0$ 和 $\alpha = 1$ 的效果都较差。当模型由 DNN 预测结果 $\hat{y}$ 和聚类结果 $\tilde{y}$ 共同的监督训练时，即 α 设置在 0.4 至 0.6 时，模型能达到较好的效果。这一结果指出，DNN 预测结果和聚类结果抛弃掉任何一个都会导致较差的结果。需要结合两者共同的监督模型的训练。



(a) (b)

图 8: 从 0.0 到 1.0 不同的 α 取值下测试准确度的变化。(a) 20% 随机噪声 CIFAR10 数据集实验结果。(b) 40% 随机噪声 FASHION-MNIST 数据集实验结果。

3.4 总结

在本章中，我们提出一个新的噪声标签学习方法 S^{2}LC (Spatial Structure mining and dynamic Label Correction)。S^{2}LC 可以被分为两个组成部分，即样本选择和标签纠正。具体地，我们的样本选择策略是基于 DNN 预测结果和训练集观测标签的一致性。经过样本选择后剩余的样本将会利用基于混合模型的 K 均值聚类结果进行标签纠正。这两个部分和 DNN 的训练交替地进行。同时，为了充分的利用 DNN 对于干净样本的学习能力，DNN 的训练将由纠正后的标签和前一轮迭代的 DNN 预测结果共同监督。为了使得样本在特征空间满足同类近，异类远的分布，提升聚类算法的准确度，我们优化 $(K+1)$ 元组损失进行判别性特征学习。我们进行了一系列的实验，实验结果证明了我们的样本选择，标签纠正和判别性特征学习的有效性。对比实验结果证明了我们的方法能超过其他的相关方法。

4 基于分布纠正的噪声标签学习算法

4.1 引言

在噪声标签学习问题中，相关工作的理论推导或是实验虚拟数据集大多构建在类别依赖噪声上。其中类别依赖噪声在章节 2.1中进行了详细说明。然而实例依赖噪声是在实际生活中更常见，也是更为困难的一种噪声模式。我们需要寻找一个噪声标签学习方法能对所有的噪声模式都具有鲁棒性。因此我们从分布的角度解决噪声标签问题。噪声标签学习问题可以归结为分布纠正问题。即将噪声分布纠正为潜在的训练集真实分布。我们使用样本加权的方式经验上的纠正噪声分布，并给出理论证明我们的权重因子能够将噪声分布纠正为潜在的真实分布。我们的权重系数可以通过混合密度估计得出。DisCo 模型采用课程学习框架交替的利用教师模型进行样本权重的更新和学生分类器的训练。不同于 CurriculumNet[109]，DisCo 的课程设计的目标不是衡量样本的难易程度，而是由教师模型输出权重以将训练集的噪声分布纠正为真实分布。另一个将课程学习用于噪声标签学习的经典工作是 MentorNet[93]。MentorNet 需要一个经过清洗的额外数据集训练学生模型，使得学生网络能输出衡量样本标签干净程度的权重。然而经过清洗的额外数据集在实际应用中不一定能够获得。并且对于噪声数据，MentorNet 将赋予小权重，从而这些样本的信息并未得到有效利用。与 MentorNet 不同，DisCo 不需要经过清洗的额外数据集，同时我们提出推广式 Mixup 使得噪声样本也能得到有效利用。

本章提出一种对包括实例依赖噪声在内的多种噪声模式鲁棒的噪声标签学习方法 DisCo（**Dis**tribution **Co**rrection）。DisCo 可以有效的纠正由噪声标签导致的后验分布偏移问题，同时使用数据扩增和重采样方法增强模型的泛化能力。DisCo 算法包含三个部分。第一个部分是样本选择。我们通过样本与决策边界的靠近程度判断它们是否为置信样本。第二个部分是样本加权。样本权重通过一个用选择的置信样本训练的原型学习模块计算。第三个部分是数据插值和扩增。不置信样本的标签会首先用原型分类器的分类结果进行标签纠正。接下来会通过任意两张图像的线性组合产生虚拟样本并利用这些样本使用加权交叉熵进行训练。

我们的工作的主要贡献总结如下：

- 我们提出了一个新的噪声学习方法 DisCo。DisCo 可以通过样本加权纠正噪

声标签导致的分布偏差。我们在理论上证明了我们的重要性加权策略的有效性。样本权重可以通过一个用所选择的置信样本训练的原型模块计算得出。

- 为了进一步地提高预测模型的泛化能力，我们首先使用基于原型的预测结果纠正不置信样本的标签，接下来通过对置信样本和标签纠正样本进行线性插值生成虚拟样本。这可以视作一种推广式的数据增广方式。
- 我们在虚拟噪声数据集和大规模现实噪声数据集上进行了一系列的实验。实验结果证明 DisCo 超越了现有的相关深度学习方法。

我们将先总结 DisCo 算法的组成部分和训练过程。接下来将分别介绍 DisCo 算法的样本选择，重要性加权和推广式 Mixup。最后将在课程学习的框架下介绍 DisCo 模型的训练过程。

4.2 DisCo 算法描述

Fig.9 展示了 DisCo 的算法框架。DisCo 算法由三个部分组成，分别是，1) 样本选择，2) 样本加权，和 3) 推广式 Mixup。在第一个部分，我们将原始数据集分成置信集和不置信集。在第二个部分，我们将估计增广训练集中样本的权重。在第三部分，我们将从纠正后的分布中采样样本以扩增训练数据集。DisCo 的训练过程采用课程学习框架，在课程设计阶段，我们使用选择的置信样本训练教师模型得到各个样本的权重系数。在课程学习阶段，我们使用推广式 Mixup 扩增训练集，并使用加权交叉熵训练学生分类器。DisCo 算法伪代码见算法 3。

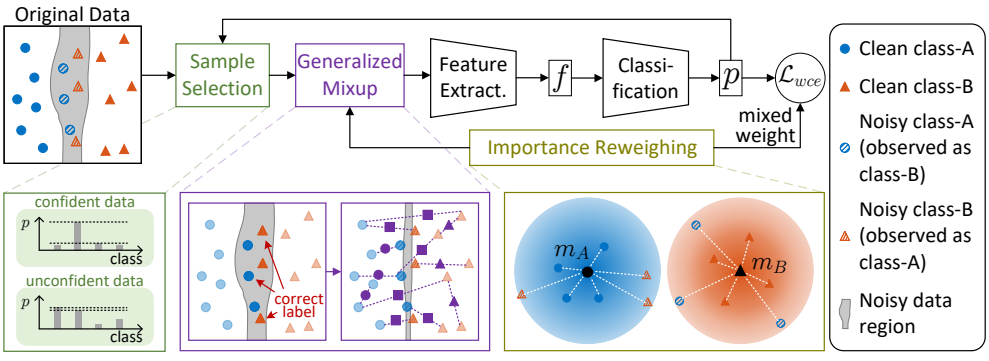


图 9: DisCo 框架

Algorithm 3 DisCo 伪代码

输入: 含有噪声标签的训练数据集 $\mathcal{D}^o = \{(x_i, y_i^o)\}_{i=1}^n$; 训练迭代次数 E; 批大小 n_b, 学习率 ϵ.

参数: 特征提取器参数 θ, 分类器参数 ϕ, 原型学习模块参数 ψ.

Output: 特征提取器 $G(\cdot;\theta)$, 分类器 $F(\cdot;\phi)$, 原型学习模块 $T(\cdot;\psi)$.

预训练 $G(\cdot;\theta)$, $F(\cdot;\phi)$, 和 $T(\cdot;\psi)$

for $t = 1$ to E **do**

  从 $\mathcal{D}^0$ 中采样 mini-batch $\mathcal{D}_b^o = \{(x_i, y_i^o)\}_{i=1}^{n_b}$;

  利用 4.1 将 $\mathcal{D}_b^o$ 分成置信集 $\mathcal{D}_b^{cf} = \{(x_i, y_i^o)\}_{i=1}^{b_n^{cf}}$ 和非置信集 $\mathcal{D}_b^{uf} = \{(x_i, y_i^o)\}_{i=1}^{b-N^{uf}}$;

  在置信集 $\{(x_i, \hat{y}_i)\}_{i=1}^{b_n^c}$ 利用 4.10 更新 ψ，其中 $\hat{y}_i = F[G(x_i;\theta);\phi]$;

  生成虚拟样本 $\mathcal{D}_b^{mix} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{n_b}$。利用 2.23 和 4.13估计样本权重;

  利用 4.14和 $\mathcal{D}_b^{mix}$ 更新 θ 和 ϕ;

end for

4.2.1 基于决策边界的样本选择

记 $\mathcal{D}^o = \{(x_i, y_i^o)\}_{i=1}^N$ 为从联合分布 $Q(X, Y^o)$ 中采样的噪声训练数据集。$\mathcal{D}^T = \{(x_i, y_i)_{i=1}^N\}$ 为从真实未知训练分布中采样的数据集。其中 $y_i^o, y_i \in \mathcal{Y} = \{1, ..., K\}$ 代表 K 个类别。如果 $y_i^o = y_i$，则 x_i 为干净样本，反之则为噪声样本。针对分类任务，我们记 x_i 的预测概率为 $P(Y|x_i) \in \mathbb{R}^K$。

从图 1(a)中可以看出，噪声样本通常分布在决策边界附近。我们可以通过分类器的预测概率来估计样本与决策边界的靠近程度。通过这种估计，我们可以判断样本的分类器预测是否与其真实标签一致。

基于上述分析，我们提出以下样本选择准则，将原始数据集分成置信子集 $\mathcal{D}^{cf}$ 和不置信子集 $\mathcal{D}^{uf}$。

$$x_i \in \begin{cases} \mathcal{D}^{cf}, P(Y = k_i|x_i) - \max\limits_{c \in \mathcal{Y}, c \neq k_i x} P(Y = c|x_i) > \theta \\ \mathcal{D}^{uf}, otherwise \end{cases} \tag{4.1}$$

其中 $i \in \{1, \cdots, N\}$, $k_i = \arg\max_{c \in \mathcal{Y}} P(Y = c|x_i)$。$\theta$ 是置信度阈值。对于一个样本

x_i, 如果其预测概率向量的*最大两个值的差*小于等于 θ, 我们将其视作不置信样本。

一种朴素的置信样本选择方法是将样本观测标签预测概率低的样本视作不置信样本。对比起这种朴素的方法，我们基于式 4.1的样本选择准则能够准确的识别出更多的不置信样本。具体地说，在多分类任务场景下，当分类器将某个样本在*多个类*之间混淆，则预测概率向量中的所有元素（包括最大的元素）都会接近于 0。此时，朴素的样本选择准则和我们的准则都可以将其判定为不置信样本。然而，当分类器仅将某个样本在*较少类*之间混淆时，样本观测标签的预测概率将会较高。在这种情况下，朴素准则很有可能将其视作置信样本，此时我们的准则仍然能将其正确判定为不置信样本。

4.2.2 基于分布纠正的重要性加权

分类任务的目的是学习一个条件分布 $\hat{P}(Y|X)$，使其能够近似真实的后验分布 $P(Y|X)$。利用交叉熵度量两个分布间的差异，我们可以得到需要最小化的期望风险

$$\arg\min \mathbb{E}_{(x,y)\sim P(X,Y)} P(y|x) \log \hat{P}(y|x). \tag{4.2}$$

然而，$P(X,Y)$ 和 $P(Y|X)$ 均是未知的。我们能够获取的只有来自于噪声联合分布 $Q(X|y^o)$ 的噪声数据集 $\mathcal{D}^o$。我们提出重加权训练样本以纠正噪声分布来取代直接最小化 $\hat{P}(Y|X)$ 和 $Q(\hat{Y}|X)$ 之间的差异，即

$$-\frac{1}{N}\sum\nolimits_{i=1}^{N} w_i Q(y_i|x_i) \log \hat{P}(y_i|x_i) \tag{4.3}$$

我们可以利用权重因子 $w_i = \frac{P(x_i|y_i)}{Q(x_i|y_i)}$ 经验上的纠正噪声分布下的交叉熵损失。这个性质将在下述命题中阐述。

结论 4.1. *对于噪声分布 $Q(X,Y)$ 和真实分布 $P(X,Y)$，假设 $P(X) = Q(X)$, $P(Y) = Q(Y)$, 重要性权重因子定义为:*

$$w_i = \frac{P(x_i|y_i)}{Q(x_i|y_i)} \tag{4.4}$$

则噪声分布 $Q(X,Y)$ 下的加权交叉熵为真实分布下的交叉熵，即，

$$-\frac{1}{N}\sum_{i=1}^{N} w_i Q(y_i|x_i) \log \hat{P}(y_i|x_i) = -\frac{1}{N}\sum_{i=1}^{N} P(y_i|x_i) \log \hat{P}(y_i|x_i) \tag{4.5}$$

对于存在标签噪声的分类任务，我们通常用 $P(X) = Q(X)$。又因为在现实生活中，数据集的标准错误通常不会改变类别比例，即，少数类不会因为标注错误而成为多数类。因此我们可以进一步的假设 $P(Y) = Q(Y)$。由于我们所能获取的仅仅是来自噪声分布 Q 的样本，因此我们假设 $P(X) = Q(X) = \frac{1}{N}, Q(y = y_i|x_i) = 1$ 和 $Q(y \neq y_i|x_i) = 0$。则我们可以经验上的将 $Q(x_i|y_i)$ 估计为 $\frac{1}{n_{y_i}}$。基于上述分析，权重因子可以被经验上的估计为

$$w_i = P(x_i|y_i)n_{y_i}, \tag{4.6}$$

其中 $P(x_i|y_i)$ 可以通过混合密度估计获得。

我们可以利用一个教师模型 $T(\cdot;\psi)$ 估计式 4.6。近期原型学习 [124, 125, 126] 被广泛的应用于混合密度估计。原型学习即将原始数据映射到特征空间 $\mathcal{T}$ 中，并在 $\mathcal{T}$ 中学习代表 K 个混合分布的原型 $m_k, k = 1, ..., K$。我们利用一个原型学习模块作为教师网络进行密度估计从而在训练过程中动态的计算样本权重。

原型学习的目标函数可以分为两个部分 [126]。第一项为最大化似然函数，由式 2.15导出，即，

$$l_1 = \sum_{i=1}^{N}\sum_{k=1}^{K} \chi_i(t_i - m_k)^2, \tag{4.7}$$

其中 $t_i = T(G(x_i;\theta);\psi)$，$m_k(k = 1, 2, ..., K)$ 为学习得到的原型，χ_i 为示性函数。假设所有的混合分布具有相同的权重，我们可以导出类别条件分布为

$$\hat{p}(y = k|G(x_i;\theta)) = \frac{\exp(-\gamma\|t_i - m_k\|_2^2)}{\sum_{k'=1}^{K}\exp(-\gamma\|t_i - m'_k\|_2^2)}, \tag{4.8}$$

其中 γ 为控制概率平滑程度的超参数。

为了让每个样本靠近其所属的类原型，第二项最大化置信样本集的样本在观测标签上的条件概率，即，最小化其负对数

$$l_2 = -\sum_{i=1}^{N}\log\hat{p}(y_i|G(x_i;\theta)). \tag{4.9}$$

原型学习的总损失函数为

$$l_{pl} = l_1 + l_2 \tag{4.10}$$

基于学习到的类原型 $m_k(k=1,...,K)$，我们可以得到样本与类原型之间的距离矩阵 $D \in \mathbb{R}^{N\times K}$，并获得 $P(x|y)$ 的估计：

$$\begin{aligned} D_{i,k} &= \|t_i - m_k\|_2^2 \\ P(x_i|y=k) &= c \cdot \exp(-\gamma D_{i,k}) \end{aligned} \tag{4.11}$$

4.2.3 基于推广式 Mixup 的数据扩增

即使我们通过赋予噪声样本小的权重经验上的矫正了噪声分布，噪声样本包含的信息却被丢弃了。因此这种纠正依然是有偏差的。为了解决这个问题，我们提出基于推广式 Mixup 的数据增广方法。这种方法通过数据插值以获取从纠正后的分布中采样的虚拟样本用以扩增数据集。

首先我们需要进一步的纠正重加权后的样本分布。在章节 4.2.1中，我们获得了不置信数据集 $\mathcal{D}^{uf}$，大多数属于不置信数据集的样本都是噪声样本，即意味着它们被赋予了较小的权重。为了充分利用这些样本所包含的信息，我们通过基于原型的预测标签纠正不置信集的样本标签，即，

$$y_i^p = \arg\max_{k\in\mathcal{Y}} \hat{p}(y=k|f_i). \tag{4.12}$$

经过上述标签纠正操作，我们在所获得的标签纠正数据集 $\mathcal{D}^{crt} = \{(x_i, y_i^c)\}_{i=1}^N = \{(x_i, y_i^p)\}_{i\in\mathcal{D}^{uf}} \cup (x_i, y_i^o)_{i\in\mathcal{D}^{cf}}$ 上计算权重。对于 $\mathcal{D}^{uf}$ 中的样本，如果在我们的重加权策略下，纠正后的标签大概率是真实的，则其会被赋予一个大的权重，反之其会仍然被赋予小权重，并可以被视作离群点。

接下来，我们将介绍推广式 Mixup 策略。基于原始的 Mixup[114]，虚拟样本由 $\mathcal{D}^{crt}$ 中随机两个样本 (x_i, y_i^c) 和 (x_j, y_j^c) 的线性组合（式 2.23）生成。对于生成的虚拟样本 $\tilde{x}_{ij}$，我们利用 (x_i, y_i^c) 和 (x_j, y_j^c) 对应的权重 w_i 和 w_j 的线性组合赋予其权重。

$$\tilde{w}_{ij} = \lambda w_i + (1-\lambda) w_j \tag{4.13}$$

4.3 课程学习角度模型分析

DisCo 也可以视作一个课程学习的框架。我们将课程设计的目标由挑选简单样本用于学生网络训练，转换为计算样本权重用于纠正噪声分布。在课程学习阶段，

我们使用推广式 Mixup 用于进一步的纠正噪声分布并扩增数据集，进一步的增强模型的泛化性能。

4.3.1 课程设计

我们的课程设计由两部分组成：1. 样本选择；2. 样本权重估计。对于样本选择，我们的策略是基于在现实生活中，标记错误的样本大多是易混淆样本（图 2），即噪声样本在特征空间上大多分布在决策边界附近。我们使用学生网络分类器对样本的预测概率估计样本与决策边界的靠近程度，进而将离决策边界较远的样本选为置信样本。对于样本权重估计，我们给出理论证明我们的权重因子可以经验上的纠正噪声分布。同时这个权重因子可以由混合密度估计得出。在神经网络的框架上，我们采用一个 MLP 原型学习模块来实现混合密度估计并得到权重因子。

4.3.2 课程学习

在课程学习阶段，我们使用加权交叉熵损失函数和推广式 Mixup 生成的虚拟数据训练学生网络。对于生成的虚拟样本 $\tilde{x}_{ij}$，加权交叉熵损失为：

$$l_w = \tilde{w}_{ij}[l_{ce}(G(F(x_i)), y_i^c) + l_{ce}(G(F(x_j), y_j^c))] \tag{4.14}$$

DisCo 模型的训练过程分为三个阶段。

1. **预热训练：** Sanchez 和 Ortego 等人 [106] 发现深度神经网络在拟合噪声样本之前，会先拟合正确标注的干净样本。在预热训练阶段，我们将使用原始带噪声数据集和交叉熵损失训练学生网络分类器。使学生分类器初步学习到决策边界。此阶段仅更新学生分类器和特征提取器参数：

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \sum_{i=1}^{N} l_{ce}(y_i^o, x_i) \tag{4.15}$$

2. **教师模型训练：** 在教师模型训练阶段，教师模型将使用基于式 4.1选出的置信样本集 $\mathcal{D}^{cf}$ 及其对应的分类器预测结果训练教师模型。此阶段在更新学生分类器，特征提取器参数的基础上额外更新教师模型的参数。

$$\psi^* = \arg\min_{\psi} \sum_{x_i \in \mathcal{D}^{cf}} l_{pl}(\hat{y}_i, G(x_i)) \tag{4.16}$$

3. **DisCo 训练：**在 DisCo 模型的训练过程中，学生分类器和教师模型交替优化。教师模型动态的更新样本权重，学生分类器使用加权交叉熵和推广式 Mixup 虚拟样本进行参数更新。

$$\begin{aligned} \theta^*, \phi^* &= \arg\min_{\theta,\phi} l_w \\ \psi^* &= \arg\min_{\psi} \sum_{x_i \in \mathcal{D}^{cf}} l_{pl}(\hat{y}_i, G(x_i)) \\ w^* &= g_m(T(G(x_i); \psi^*)) \end{aligned} \tag{4.17}$$

其中 g_m 为将基于式 4.6和 4.11的权重转换函数。

4.4 实验与分析

我们在虚拟数据集和大型真实数据集上进行实验检验 DisCo 模型的效果。我们将在本章节中介绍实验细节和实验结果与分析。

4.4.1 比较方法

我们将 DisCo 与以下代表方法进行对比。分别为，交叉熵损失（CE），Forward[62]，CRUST[57]，Boostrapping（BS）[53]，PLC[127]，IDN[58]，GCE[54]，SCE[55]，LRT[59]，MetaCleaner[52]，CleanNet[60] 和 SMP[56]。所有方法的超参数选择都基于验证集的表现。

4.4.2 实验细节

对于噪声 MNIST 数据集，我们使用和 [58] 中一致的特征提取器和分类器网络结构。对于噪声 CIFAR-10 数据集，我们使用 PreAct ResNet-32[123] 作为特征提取器和分类器。对于两个虚拟数据集，均使用包含一层全连接隐藏层的 MLP 作为原型学习模块。初始学习率均设定为 0.01，在第 40 次和第 80 次迭代进行减半的学习率衰减。网络训练优化器和数据预处理均与 [67] 中一致。

对于 Clothing1M 和 Food101N 数据集，我们均使用 ResNet-50 作为特征提取器和分类器。初始学习率均设定为 0.001。原型学习模块的结构和虚拟数据集实验中的设定一致。

4.4.3 对比实验结果与分析

使用存在噪声标签的数据集训练 DNN 通常会出现过拟合现象，即训练最后一次迭代的测试准确度会远低于验证集效果最好的一次迭代对应的测试准确度。基于这个现象，我们记录了验证集效果最好的一次迭代的测试结果（记为 **BEST**）和最后一次迭代的测试结果（记为 **LAST**）。如果这两个结果相差无几，则证明模型可以避免过拟合。

噪声 MNIST 数据集实验：从表 7中可以看出，DisCo 在所有的噪声比例下在验证集最优迭代和最后一次迭代下均取得了最优的测试集准确度。在 DisCo 的训练过程中，模型测试准确度即使在训练后期也一直保持稳定。这证明我们的模型对噪声标签鲁棒，并且能够有效避免过拟合。在对比方法中，Forward[62] 的测试准确度比 CE 基线还低。出现这个现象的原因可能是 Forward[62] 方法的构建是基于类别依赖噪声假设的，而我们所模拟的噪声模式是更为现实且更具挑战性的实例依赖噪声。对比起其他专门用于解决实例依赖噪声的方法，即 IDN[58] 和 PLC[127]，DisCo 分别将测试集准确度提高了 6% 和 3%。

噪声 CIFAR10 数据集实验：表 8中的结果展示了我们的方法整体地超过了其他的对比方法。虽然部分方法具备避免过拟合的能力，但它们的测试准确度却仅仅和 CE 基线持平，甚至低于 CE。比如 GCE[54]，虽然它的测试准确度在整个训练过程中都保持平稳，但是其测试准确度却比 CE 低将近 2%。这些方法大多将注意力放在增强模型的泛化能力上，却忽视了模型的拟合能力。CRUST[57] 和 DisCo 均采用了样本选择的框架。然而 DisCo 在所有噪声设定上总体都超越了 CRUST[57]。CRUST 和 DisCo 的区别在于，CRUST 用所选择的样本在下一轮迭代中重新训练分类器，而 DisCo 则用所选择的样本训练一个另外的原型学习模块。对比起用所选择的样本重新训练原本的分类器，训练另一个模型可以有效避免样本选择偏差带来的累积误差。

真实噪声数据集实验：表 9展示了在 Clothing1M 和 Food101N 上的实验结果。我们可以看出 DisCo 在两个数据集中均取得了最优的效果。具体地，DisCo 在 Food101N 和 Clothing1M 上对比起 CE 分别获得了 4.56% 和 5.20% 的提升。对于 Clothing1M，DisCo 大幅超出鲁棒性损失函数方法（即，Forward，GCE 和 SCE）。导致鲁棒性损失函数效果欠佳的原因可能是所有的损失函数的目标都是学习一个能够近似训

表 7: 噪声 MNIST 数据集实验结果

method		MNIST			
noise rate (%)		0.1	0.2	0.3	0.4
CE	BEST	96.61	93.82	90.65	85.37
	LAST	95.98	90.18	80.66	67.87
Forward [62]	BEST	96.38	92.96	89.31	83.05
	LAST	94.78	86.54	76.55	65.68
CRUST [57]	BEST	96.22	94.93	93.82	88.19
	LAST	95.8	94.64	92.71	86.39
BS [53]	BEST	96.61	93.82	91.65	88.86
	LAST	96.12	93.63	90.12	81.55
PLC [127]	BEST	97.21	95.69	93.82	91.27
	LAST	97.08	95.98	93.53	90.12
IDN [58]	BEST	96.66	94.74	92.31	90.1
	LAST	96.06	94.38	91.95	89.16
GCE [54]	BEST	96.59	93.95	90.40	86.52
	LAST	96.26	91.49	88.66	84.52
DisCo	BEST	**98.10/97.44**	**97.69/97.06**	**97.27/96.70**	**96.13/95.64**
	LAST	**98.00/97.28**	**97.49/96.92**	**97.01/96.32**	**95.74/95.45**

表 8: 噪声 CIFAR10 数据集实验结果

method		CIFAR10			
noise rate (%)		0.1	0.2	0.3	0.4
CE	BEST	86.24	83.71	81.2	78.4
	LAST	85.34	82.53	77.5	74.2
Forward [62]	BEST	88.40	85.32	82.03	79.06
	LAST	85.74	82.28	78.44	72.34
CRUST [57]	BEST	88.40	85.02	81.39	78.03
	LAST	88.33	84.57	80.32	77.74
BS [53]	BEST	87.2	83.42	78.5	78.5
	LAST	86.71	82.7	74	74
PLC [127]	BEST	86.02	83.96	80.02	78.45
	LAST	85.39	83.39	80.70	78.03
IDN [58]	BEST	88.27	85.12	83.24	80.98
	LAST	87.93	83.83	82.68	80.56
GCE [54]	BEST	85.67	83.64	81.12	76.79
	LAST	85.7	82.96	80.39	76.77
DisCo	BEST	**88.19/88.17**	**86.87/86.92**	**84.78/84.89**	**81.96/82.02**
	LAST	**87.32/87.52**	**86.07/86.92**	**84.00/84.09**	**81.00/81.30**

练集条件分布的近似分布，然而当训练集条件分布被噪声标签影响的时候，使用这些损失函数进行训练实际上得到的都是一个有偏差的近似分布。同时，DisCo 在 Clothing1M 数据集上也超过了专门为实例依赖噪声设计的算法 PLC[127]。在 Food101N 数据集上，DisCo 超过了 CleanNet[60]。CleanNet[60] 的训练需要额外的干净数据集而 DisCo 不需要。

表 9: 真实噪声数据集实验结果

(a) Clothing1M

Method	Accuracy(%)
CE	69.04
Forward[62]	69.84
GCE[54]	69.75
SCE[55]	71.02
LRT[59]	71.74
MetaCleaner[52]	72.5
PLC[127]	74.02
DisCo	**74.24/74.27**

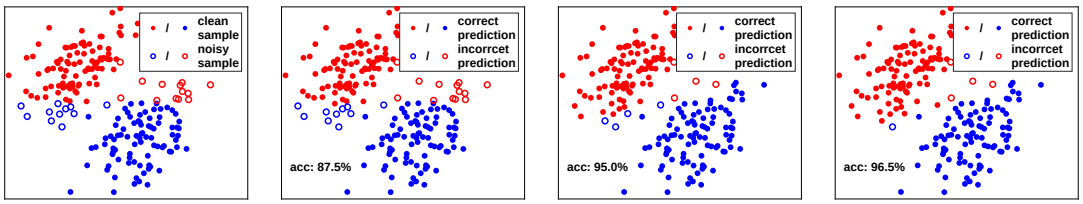
(b) Food101N

Method	Accuracy(%)
CE	81.03
CleanNet[60]	83.95
GCE[54]	84.12
PLC[127]	85.28
MetaCleaner[52]	85.05
SMP[56]	85.11
DisCo	**85.59/85.22**

4.4.4 模型分析

置信样本选择：我们首先在虚拟数据集上进行了实验证明 DisCo 可以准确地分类决策边界附近的样本。我们首先利用两个不同的高斯分布生成两个类别的虚拟数据。接下来依照 [58]，我们错误标注决策边界附近的样本构造噪声数据集（图 10(a)。从图 10(b)中可以看出，仅使用 CE 进行训练无法使分类器准确分类决策边界附近的样本。图 10(c)和图 10(d)展示了 DisCo 中全连接分类器（FCC）和原型学习分类器（PL）的分类结果。我们可以看出 FCC 将绝大多数决策边界附近的样本分类正

确，而 PL 的表现会更加好。这验证了 DisCo 具备很强的对决策边界附近样本的分类能力。



(a) (b) (c) (d)

图 10: 在虚拟数据集上的实验结果证明了 DisCo 能够有效的分类决策边界附近的样本。(a) 生成的虚拟噪声数据集。(b) CE 的预测结果。(c) DisCo 中 FCC 的预测结果。(d) DisCo 中 PL 的预测结果。

我们使用标签纯净度指标衡量我们的置信样本选择策略的效果。

$$pure\ ratio \triangleq \frac{\{|y_i^p = y_i | x_i \in \mathcal{D}^{cf}\}}{\mathcal{D}} \tag{4.18}$$

我们将 DisCo 与另一个使用了样本选择框架的方法 Co-teaching 进行对比。Co-teaching 选择损失较小的一部分样本与这部分样本的观测标签训练另一个分类器。Co-teaching 和 DisCo 的目标都是选择出干净的样本集用于训练另一个分类模型。图 12展示了 DisCo 和 Co-teaching 在训练过程中选择出的数据集的标签纯净度。我们可以看出 DisCo 选择出的数据集的标签纯净度远超 Co-teaching 。事实上，即使在训练的初期，DisCo 选择出的数据集的标签纯净度就已经接近 100%。这意味着 DisCo 能够提供近乎完全干净的数据集用于训练原型学习模块。因此 DisCo 的原型学习模块能够准确的给予一个精确的权重估计。

重要性样本加权： 图 13(a)和 13(b)分别展示了 DisCo 方法赋予干净样本和噪声样本的权重。我们可以看出干净样本被赋予了高的权重（展示为红色），而噪声样本被赋予了小的权重（展示为蓝色）。我们可以得出结论 DisCo 可以准确的区分干净样本和噪声样本，并使模型更注重学习干净样本中的标签-样本映射信息，同时避免噪声样本带来的负面影响。

推广式 Mixup： 图 14展示了 DisCo 中的推广式 Mixup 生成的样本的联合分布 $P(\tilde{X}, \tilde{Y})$，其中 $\tilde{Y} \in [0, 1]$ 为 Mixup 样本的软标签。红色代表类别 1，蓝色代表类别

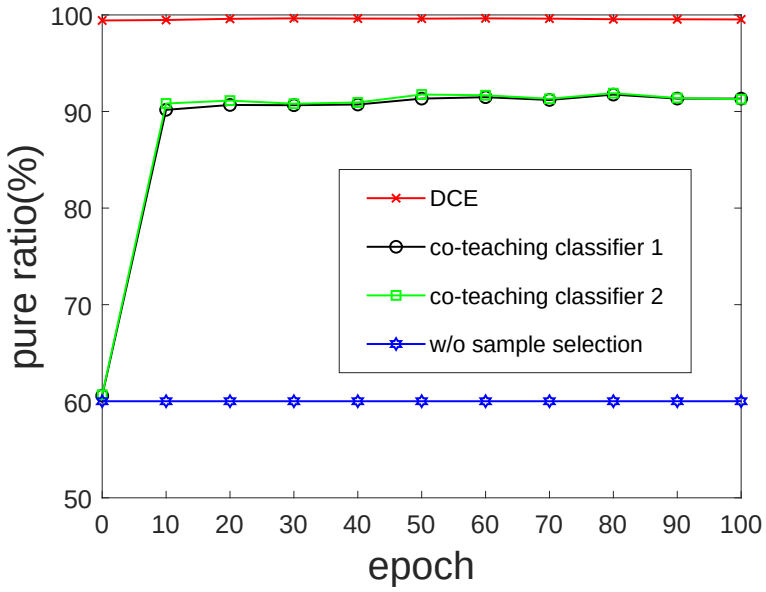
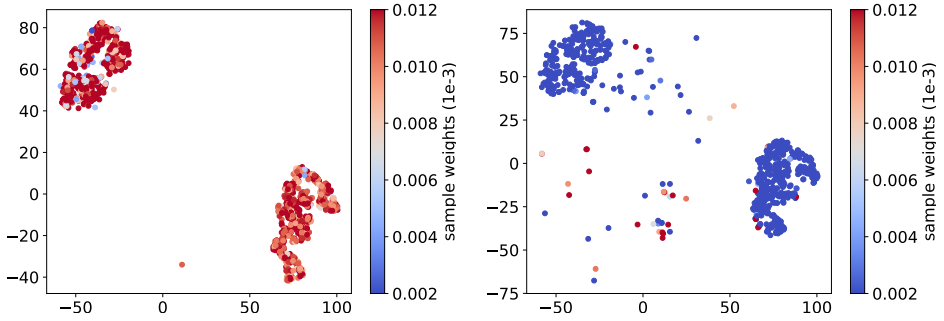


图 11: 被选择样本的标签纯净度 vs. MNIST 数据集实验迭代次数。

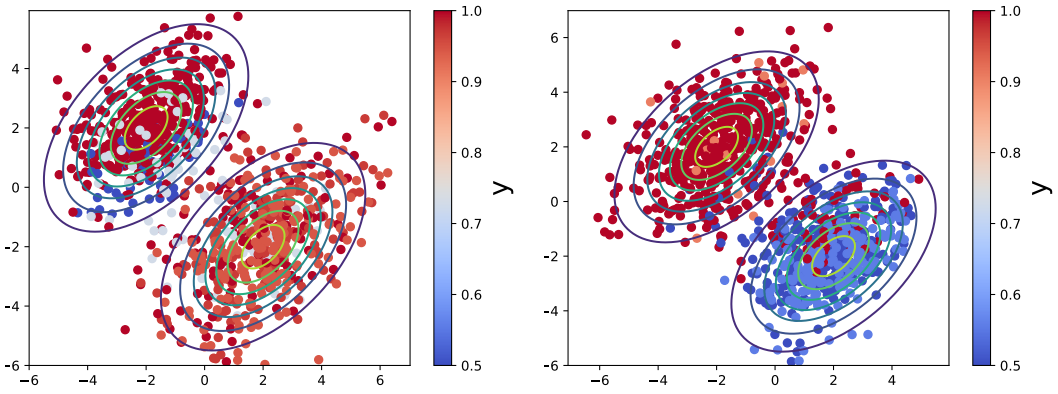
图 12: Label pure ratio of selected samples vs. number of epochs on MNIST dataset.



(a) (b)

图 13: DisCo 方法在 40%MNIST 噪声数据集上的特征可视化。(a) 干净样本权重热力图。(b) 噪声样本权重热力图。DisCo 给噪声样本赋予了小权重。

0。由于我们所使用的虚拟数据集是由两个不同的高斯分布生成的（可视化为 14中的等高线），来自真实训练分布的数据应该分布在两个高斯分布的等高线附近，并且有可判别为两类的颜色。然而如图 14(a)所示，用经典的 Mixup[114] 生成的样本并未明显的在两个高斯分布的附近分为两类（绝大多数的样本都是红色的）。而 DisCo 中采用的推广式 Mixup 方法生成的样本分布在两个高斯分布的等高线附近并且从颜色上可以清晰的分为两类。这证明了 DisCo 中的推广式 Mixup 可以更精准的近似潜在的训练分布并从中采样样本已扩充训练集。



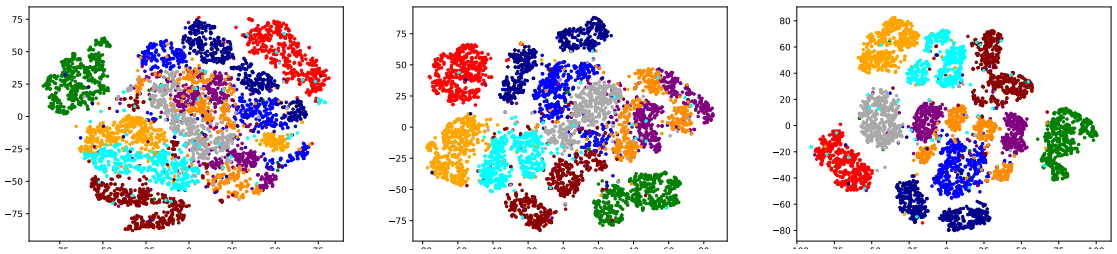
(a) Mixup contour (b) DisCo contour

图 14: 生成的 Mixup 样本及其对应的软标签的联合分布。等高线代表真实的潜在训练分布。红色代表类别 1，蓝色代表类别 0。（a）使用经典 Mixup[114] 和样本加权生成的样本。（b）DisCo 生成的样本。

消融实验：我们在噪声 MNIST 数据集进行了消融实验以展示 DisCo 模型中各组成部分的有效性。我们将模型最后一轮迭代的测试集结果记录在表 10。我们可以看出推广式 Mixup 和样本重要性加权都能提高模型的表现。而 DisCo 在所有的噪声比例上都取得了最优的效果。实验结果显示 DisCo 模型中的所有部分都是对训练有提升的。同时我们将样本在特征空间上的分布可视化，如图 15所示。从图 15(a)中可以看出，当仅使用 CE 进行训练时，噪声标签会导致样本在特征空间在不同类之间混淆。图 15(b)展示出经典的 Mixup[114] 具有使样本特征更具类别判别度的能力，而 DisCo 能够更进一步的使同类样本特征分布紧致，异类样本特征分布分散，如图 15(c)所示。

表 10: Ablation study

noise rate (%)	CE	Mixup	Reweighing	DisCo
0.1	95.98	97.44	97.23	**98.00/97.28**
0.2	90.18	95.58	95.60	**97.49/96.92**
0.3	80.66	92.91	94.04	**97.01/96.32**
0.4	67.87	89.20	91.54	**95.74/95.45**



(a) CE (b) Mixup+CE (c) DisCo

图 15: 在 40% 的噪声 MNIST 数据集上的特征可视化。(a) 使用 CE 训练。(b) 使用 CE 和 Mixup 进行训练。(c) 使用 DisCo 进行训练。Mixup 和样本加权有效的在噪声标签的影响下将不同类别的特征分隔开。

4.5 本章总结

在这一章中，我们提出了一种基于噪声分布纠正的噪声标签学习方法 DisCo。DisCo 分为三个重要组成部分，分别是样本选择，重要性加权和推广式 Mixup 数据扩增。我们首先介绍了这三个组成部分。接下来，我们从课程学习的角度介绍了 DisCo 的训练过程。样本选择和重要性加权可以视作课程设计的组成部分，课程学习通过加权交叉熵损失和推广式 Mixup 生成的虚拟样本进行。我们在两个虚拟噪声数据集和两个大规模真实噪声数据集上的实验验证了本文提出的 DisCo 方法的效果。具体地说，在与其他方法的对比实验中，DisCo 不管是在虚拟噪声数据集还是真实数据集上都整体地超过了其他的对比方法。我们也通过多组实验证明了 DisCo 方法中三个组成部分，即样本选择，重要性加权和推广式 Mixup 数据扩增的有效性。

5 总结

随着深度学习的发展，深度卷积神经网络在很多的计算机视觉领域的各种应用上都取得了巨大的成功，比如图像分类，物体检测和实例分割等。然而训练深度卷积神经网络需要大量的有准确标注的训练数据。在实际应用中，收集大量的带有干净标注的数据集是代价十分昂贵且花费时间的。比如医学领域的图像和精细划分图像数据集。由于收集干净数据集较为困难，因此许多专家学者转而收集来自众包平台或者搜索引擎的数据。然而这些数据通常包含着大量标注错误的样本。这些标注错误的样本被称为带有噪声标签的噪声样本。然而卷积神经网络极其容易过拟合于噪声标签。因此为了使模型的训练能够更好的适应实际应用中更常见的噪声数据集，噪声标签学习成为了一个重要的研究方向。本文针对噪声标签学习问题，分别从挖掘样本特征空间结构和分布纠正的角度提出了两个基于混合模型的方法。

第一种方法称为 S^2LC(Spatial Structure mining and dynamic Label Correction)。本方法通过基于混合模型的 K 均值聚类对噪声标签进行纠正。S^2LC 分为两个阶段，即样本选择和标签纠正。在样本选择阶段，我们将基于神经网络分类器的预测结果和聚类结果的一致性从训练集中选择出干净样本。在标签纠正阶段。对于除选择出的干净样本之外的剩余样本，我们用聚类结果纠正这些样本的标签。经过上述两个阶段，我们获得了标签经过纠正的数据集。结合自训练框架，我们使用纠正后的标签和上一轮神经网络分类器的预测结果共同的监督神经网络下一轮的训练。为了提升聚类结果的准确性，即使特征呈现同类靠近，异类远离的近似高斯混合分布的团簇状，我们额外地优化 $(K+1)$ 元组损失。我们在两个虚拟数据集，两种噪声模式和一个大规模真实数据集下进行了一系列实验。实验结果证明了我们的样本选择，标签纠正方法的有效性，和 $(C+1)$ 元组损失对模型效果的提升。与其他噪声标签学习方法的对比实验结果证明了我们的方法的优越性。

第二种方法称为 DisCo(Distribution Correction)。本方法通过基于混合模型的原型学习模块计算样本权重纠正噪声分布。DisCo 分为三个阶段，样本选择，样本加权和推广式 Mixup。在样本选择阶段，我们基于样本距离决策边界的远近程度选择置信样本，距离较远的视为置信样本。在样本加权阶段，我们使用第一阶段选择出的置信样本训练一个原型学习模块，并使用这个原型学习模块计算样本权重以纠正噪声分布。在推广式 Mixup 阶段，我们先使用原型学习模块的预测结果纠正

不置信样本的标签，接下来通过对任意两个样本的线性插值生成虚拟样本扩增训练集。我们进行了一系列的实验验证了我们方法的优越性。其中在两个虚拟数据集的噪声模式上，我们采用了更为困难的实例依赖噪声。同时我们也在两个大规模真实数据集上进行了实验。实验结果证明了我们的置信样本选择，样本加权和推广式 Mixup 各自的有效性。同时 DisCo 在所有数据集上的表现均超过了经典的和近期的噪声标签学习算法。

本文提出的 S^2LC 和 DisCo 方法是一种在带有噪声的训练集上直接进行训练的方法。在未来的研究和实际应用上，还有很多值得延拓的空间。在实际应用中，很多时候会面临着更直接的需求，即在噪声数据集中挑选出标注正确的干净样本。我们的 S^2LC 和 DisCo 方法中均包含样本选择这一阶段，因此是否能将我们的样本选择策略进行进一步的细化从而能够更加精准的挑选出干净样本是下一步的研究目标。

参考文献

[1] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 05 2017.

[4] You-Wei Luo, Chuan-Xian Ren, Dao-Qing DAI, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[5] Chuan-Xian Ren, Pengfei Ge, Dao-Qing Dai, and Hong Yan. Learning kernel for conditional moment-matching discrepancy-based image classification. *IEEE Transactions on Cybernetics*, 51(4):2006–2018, 2021.

[6] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. *Handwritten Digit Recognition with a Back-Propagation Network*, page 396 – 404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[11] Pedro H. O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *CoRR*, abs/1506.06204, 2015.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2015.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[14] Joel Akeret, Chihway Chang, Aurelien Lucchi, and Alexandre Refregier. Radio frequency interference mitigation using deep convolutional neural networks. *Astronomy and Computing*, 18:35–39, 2017.

[15] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, 09 2014.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[21] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjan Sathianathen, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

[22] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C. Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, and et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1), 07 2021.

[23] Geng-Xin Xu, Chen Liu, Jun Liu, Zhongxiang Ding, Feng Shi, Man Guo, Wei Zhao, Xiaoming Li, Ying Wei, Yaozong Gao, Chuan-Xian Ren, and Dinggang Shen. Cross-site severity assessment of covid-19 from ct images via domain adaptation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021.

[24] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[25] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds200-2011 dataset. *Advances in Water Resources - ADV WATER RESOUR*, 07 2011.

[26] W. Bi, L. Wang, J.T. Kwok, and Z. Tu. Learning to predict from crowd sourced data. *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*, pages 82–91, 01 2014.

[27] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube - a dataset for learning fine-grained action concepts from webly supervised video data, 06 2019.

[28] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: A weakly-supervised deep learning framework for learning from web data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2915–2924, 2017.

[29] Dongxu Li, Cristian Rodríguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, 10 2019.

[30] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *2015 IEEE Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 2691–2699, 2015.

[31] David F. Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.*, 33(4):275–306, 04 2010.

[32] M. Pechenizkiy, A. Tsymbal, S. Puuronen, and O. Pechenizkiy. Class noise and supervised learning in medical domains: The effect of feature extraction. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 708–713, 2006.

[33] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64, 11 2016.

[34] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 07 2019.

[35] Anders Krogh and John Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992.

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

[37] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *JMLR.org*, 2015.

[38] A. C. Atkinson and D. M. Hawkins. Identification of outliers. *Biometrics*, 37(4):860, 2018.

[39] R. J. Beckman and R. D. Cook. Outlier.....s. *Technometrics*, 25(2), 1983.

[40] Barnett, V., Lewis, T., Abeles, and Francine. Outliers in statistical data. *Phys Today*, 1979.

[41] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[42] B. S. Olkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, 2000.

[43] D A Clifton and L Tarassenko. Novelty detection in jet engine vibration spectra. *International Journal of Condition Monitoring*, 5(2):2–7, 2015.

[44] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[45] H. Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

[46] VARUN, CHANDOLA, ARINDAM, BANERJEE, VIPIN, and KUMAR. Anomaly detection: A survey. *Acm Computing Surveys*, 2009.

[47] H. Xiong, Gaurav Pandey, M. Steinbach, and Vipin Kumar. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, 18(3):304–319, 2006.

[48] Hanna Lukashevich, Stefanie Nowak, and Peter Dunker. Using one-class svm outliers detection for verification of collaboratively tagged image training sets. In *2009 IEEE International Conference on Multimedia and Expo*, pages 682–685, 2009.

[49] Fan Ma, Yu Wu, Xin Yu, and Yi Yang. Learning with noisy labels via self-reweighting from class centroids. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–11, 05 2021.

[50] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8688–8696, 2018.

[51] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Push the student to learn right: Progressive gradient correcting by meta-learner on corrupted labels. *CoRR*, abs/1902.07379, 2019.

[52] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7365–7374, 2019.

[53] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. 12 2014.

[54] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *CoRR*, abs/1805.07836, 2018.

[55] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 322–330, 2019.

[56] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5137–5146, 2019.

[57] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of neural networks against noisy labels. *CoRR*, abs/2011.07451, 2020.

[58] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[59] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris N. Metaxas, and Chao Chen. Error-bounded correction of noisy labels. *CoRR*, abs/2011.10077, 2020.

[60] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5447–5456, 04 2018.

[61] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

[62] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241, 2017.

[63] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845—869, 05 2014.

[64] J. Zhang, X. Wu, and V. S. Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):1–34, 2016.

[65] Nitika Nigam, Tanima Dutta, and Hari Gupta. Impact of noisy labels in learning techniques: A survey. pages 403–411, 01 2020.

[66] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning, 2019.

[67] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[68] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7010–7018, 2019.

[69] Peter A. Lachenbruch. Discriminant analysis when the initial samples are misclassified. *Technometrics*, 8(4):657–662, 1966.

[70] G. J. McLachlan. Asymptotic results for discriminant analysis when the initial samples are misclassified. *Technometrics*, 14(2):415–422, 1972.

[71] Peter A. Lachenbruch. Note on initial misclassification effects on the quadratic discriminant function. *Technometrics*, 21(1):129–132, 1979.

[72] Joel E. Michalek and Ram C. Tripathi. The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *Publications of the American Statistical Association*, 75(371):713–721, 2016.

[73] Yingtao Bi and Daniel R. Jeske. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Journal of Multivariate Analysis*, 101(7):1622–1637, 2010.

[74] J.S. Sánchez, F. Pla, and F.J. Ferri. Prototype selection for the nearest neighbour rule through proximity graphs. *Pattern Recognition Letters*, 18(6):507–513, 1997.

[75] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):p.257–286, 2000.

[76] Seishi Okamoto and Nobuhiro Yugami. An average-case analysis of the k-nearest neighbor classifier for noisy domains. In *International Joint Conference on Artifical Intelligence-volume*, 1997.

[77] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.

[78] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. 02 2018.

[79] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[80] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1431–1439, 2015.

[81] Alan Joseph Bekker and Jacob Goldberger. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2682–2686, 2016.

[82] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6835–6846, 2019.

[83] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *CoRR*, abs/1809.02165, 2018.

[84] Naresh Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. Cybern.*, 43(3):1146–1151, 2013.

[85] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press, 2017.

[86] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. *CoRR*, abs/2006.13554, 2020.

[87] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *CoRR*, abs/1910.03231, 2019.

[88] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W. Tsang, James T. Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *CoRR*, abs/2011.04406, 2020.

[89] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *CoRR*, abs/1903.11680, 2019.

[90] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017.

[91] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Prestopping: How does early stopping help generalization against label noise? *CoRR*, abs/1911.08059, 2019.

[92] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[93] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 07 2018.

[94] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[95] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption?, 2019.

[96] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. *CoRR*, abs/2003.02752, 2020.

[97] Yanyao Shen and Sujay Sanghavi. Iteratively learning from the best. *CoRR*, abs/1810.11874, 2018.

[98] Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. *CoRR*, abs/1905.05040, 2019.

[99] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. In H. Larochelle,

M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21382–21393. Curran Associates, Inc., 2020.

[100] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021.

[101] Markus Breunig, Hans-Peter Kriegel, Raymond Ng, and Joerg Sander. Lof: Identifying density-based local outliers. volume 29, pages 93–104, 06 2000.

[102] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Covariate Shift by Kernel Mean Matching*, pages 131–160. 2009.

[103] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. *Meta-Weight-Net: Learning an Explicit Mapping for Sample Weighting*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[104] Yifan Ding, Liqiang Wang, Deliang Fan, and Boqing Gong. A semi-supervised two-stage approach to learning from noisy labels. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1224, 2018.

[105] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[106] Eric Arazo Sanchez, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. *CoRR*, abs/1904.11238, 2019.

[107] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference*

on Machine Learning, ICML '09, page 41 – 48, New York, NY, USA, 2009. Association for Computing Machinery.

[108] M. Kumar, Ben Packer, and Daphne Koller. Self-paced learning for latent variable models. pages 1189–1197, 01 2010.

[109] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 139–154, Cham, 2018. Springer International Publishing.

[110] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[111] O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-supervised learning(chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[112] Jonathan Li and Andrew Barron. Mixture density estimation. *Advances in Neural Information Processing Systems*, 04 2000.

[113] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[114] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[115] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[116] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[117] Jun Du and Zhihua Cai. Modelling class noise with symmetric and asymmetric distributions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[118] Aditya Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption, 05 2016. arXiv preprint.

[119] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing.

[120] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[121] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics*, 52(1-2):7–21, 2010.

[122] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, and Bo An. Can cross entropy loss be robust to label noise? In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20*, 2020.

[123] K. He, X. Zhang, S. Ren, and S. Jian. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016.

[124] Brendan F. Klare and Anil K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410–1422, 2013.

[125] You-Wei Luo, Chuan-Xian Ren, Dao-Qing DAI, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[126] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3474–3482, 2018.

[127] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *ICLR*, 2021.

攻读硕士学位期间发表学术论文情况

6 致谢