基于多轮修正噪声标签的神经网络分类框架

王学刚,王玉峰
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:利用大规模的带标签数据集训练神经网络在分类任务中表现出色,但是实际使用的数据集中通常包含噪声标签从而使得分类网络的性能变差。为了克服噪声标签的不利影响,提出了一种基于多轮修正噪声标签的神经网络分类框架。该方法在每一轮修正中均更新训练的网络参数并修正当前训练数据中的噪声标签,修正后的数据集用于下一轮训练和修正。具体而言,在每一轮修正中首先利用本轮的数据集训练网络,并利用“锚点样本”的网络预测值估计数据集的标签转移矩阵;然后计算数据集的加权平均噪声率;之后结合加权平均噪声率和数据样本的训练损失值依据“小损失”原则筛选出噪声标签;最后利用标签转移矩阵和网络预测值对噪声标签进行自适应修正。经多轮修正可有效地降低数据集的噪声水平,从而使得训练出的分类网络更加准确。在多个真实数据集上的实验结果表明,该方法与现有的方案相比有较大的性能提升。

关键词:噪声标签;标签转移矩阵;加权平均噪声率;多轮修正;神经网络分类框架
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2023)08-0151-08
doi:10.3969/j.issn.1673-629X.2023.08.022

A Neural Network Classification Framework Based on Calibrating Noisy Labels in Multi-round

WANG Xue-gang, WANG Yu-feng
(School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Neural networks trained with large-scale labeled datasets have shown excellent performance in classification tasks. However, the datasets used for practical tasks often contain noisy labels, which will make the performance of the classification network worse. In order to overcome the adverse effects of noisy labels, we propose a neural network classification framework based on calibrating the noisy labels in multi-round, which updates the trained network parameters and calibrates the noisy labels in the current training data in each round of calibration, the calibrated dataset is used for the next round of training and calibration. Specifically, in each round, the predictive results of some selective anchor samples is utilized to estimate the label transition matrix of the current data, which is then used to infer the weighted average noise rate. Then, through exploiting the "small loss" principle, the noisy data samples are chosen by considering the weighted average noise rate and the trained loss value of data sample. Finally, the noisy labels are calibrated through combining the label transition matrix and predictive result of the trained network adaptively. After multiple rounds of calibration, the noisy level of the dataset is significantly reduced, based on which more accurate classification result can be trained. Experimental results on multiple real datasets show that the proposed method has a large performance improvement compared with existing schemes.

Key words: noise labels; label transition matrix; weighted average noise rate; multi-round calibration; neural network classification framework.

0 引 言

最近,基于监督学习的深度神经网络在诸多领域取得了巨大的成功[1],但是带噪声标签的训练样本,对其性能产生很大影响[2-4]。噪声标签在数据收集和标注过程中是自然存在的。雇佣专业人员来标注数据其成本非常高,因此实际中更多采用的做法是使用众包

收稿日期:2022-09-13 **修回日期**:2023-01-17
基金项目:江苏省重点研发计划(BE2020084-1);2018 年国家自然科学基金(61801240);2019 年江苏高校青蓝工程中青年学术带头人资助(QL00219001)
作者简介:王学刚(1997-),男,硕士,研究方向为弱监督学习和主动学习等;通讯作者:王玉峰(1974-),男,博士,教授,研究方向为数据科学与算法机制设计相结合的研究和应用,包括移动社交网络、众包系统、推荐系统、专家系统以及 Energy Internet 等。

平台来完成标注工作[5],如此虽降低了成本,但标注质量通常较差[6-7];另外在一些特殊领域(如医学成像),其数据标注工作非常复杂,即使专家标注也可能因为标注者的主观性或缺乏经验而产生噪声标签[8]。

对噪声标签对网络的消极影响问题,目前的解决方案包括建模标签转移模型和清除带噪数据等,其在相应的实验仿真中都展现出了一定的抗噪声性能,同时也暴露出了一定的局限性。基于建模标签转移模型的方法往往通过增加额外网络层来建模标签转移关系或利用标签转移矩阵来矫正损失函数,该类方法依赖于对标签转移关系的准确建模或标签转移矩阵的高精度估计,该工作通常存在一定的挑战;其次,基于清除带噪数据的方法是将带噪数据进行剔除或给予较小的权重,然而这样的操作不但减小了数据集的规模,并且可能忽略掉一些重要数据、破坏数据集的完整性。

针对已有方法存在的问题,该文提出了一种基于多轮修正对抗噪声标签的神经网络分类框架 MCNN。具体而言,贡献主要在以下 3 个方面:

首先,提出了一种估计数据集标签转移矩阵和噪声率的方法。与之前的工作相比,该方法不需要引入额外的人工标注成本。

其次,根据"小损失"准则提出了一种筛选带噪数据的方法,同时还提出了一种有效的方法来修正噪声标签。与之前的工作相比,提出的筛选带噪数据方法不依赖于特定的系统或网络,同时对筛选出的噪声标签进行修正,保留了数据集的完整性。

最后,在多个真实数据集上对所提出的 MCNN 进行了训练测试,表明 MCNN 能够有效地对抗噪声标签,提升模型的分类性能。

1 相关工作

近年来,缓解噪声标签对神经网络影响的方法有很多。Reed 等[9]提出了一种样本重标注的方法 Bootstrapping,使用神经网络预测值和原始标签的线性组合对全体样本进行重标注,然后进行反向传播训练网络,缺点在于在大噪声率情况下效果不佳。余等[10]也提出了一种基于重标注样本来对抗噪声标签的方法,然而局限性在于其主要适用于"二元分类"任务。

Jindal 等[11]采用增加"噪声适应层"来构建标签转移模型。通过在基本网络的 softmax 层后面增加一个 $K*K$ 维的线性约束层来建模标签转移模型,并使用正则化来惩罚线性约束层的迹以使线性约束层逼近于实际标签转移模型,然而该方法的缺点是对数据的噪声类型有一定限制。

与在基本网络后附加"噪声适应层"的方案不同,文献[12-14]将建模标签转移模型和训练分类器解耦,代表性的工作有:Patrini 等[12]提出了"前向纠正"法 F-correction 来矫正损失函数。首先,利用估计的标签转移矩阵与数据样本的网络预测值做矩阵乘积,然后,将所得结果与标签值计算交叉熵损失以此达到矫正损失的目的,但估计标签转移矩阵通常存在误差,数据量较大时存在误差累积问题。

Malach 等[15]在观察利用带噪声标签数据训练网络时发现,在整个训练过程中网络倾向于在拟合带噪声标签的数据之前先拟合带正确标签的数据,因此在网络的整个训练过程中平均损失较大的数据更大概率是带噪声标签的数据,该训练特性称作"小损失"原则。

之后利用"小损失"原则产生了一系列对抗噪声标签的方法[16-18],代表性的工作有:Han 等[16]提出了一种名为 Co-teacher 的联合学习方法,该方法同时训练两个不同的网络,两个网络都基于本网络中样本的损失值移除可能的带噪声标签数据,然后将去噪后的数据传给另一个网络做下一次迭代更新,然而该方法在训练后期存在网络"靠拢"问题,另外将含噪数据移除有可能意外地去除一些有用样本,同时假设数据集的噪声率是已知的不符合真实的任务场景。

2 MCNN 框架和组成部分

2.1 MCNN 基本模型结构

如图 1 所示,MCNN 网络框架在每轮修正中包括 4 个步骤:估计标签转移矩阵、估计数据集的噪声率、筛选可能的带噪数据和自适应修正带噪数据中的标签。注意,在每一轮修正中,MCNN 都更新所训练的 DNN 网络以及所使用的数据集。

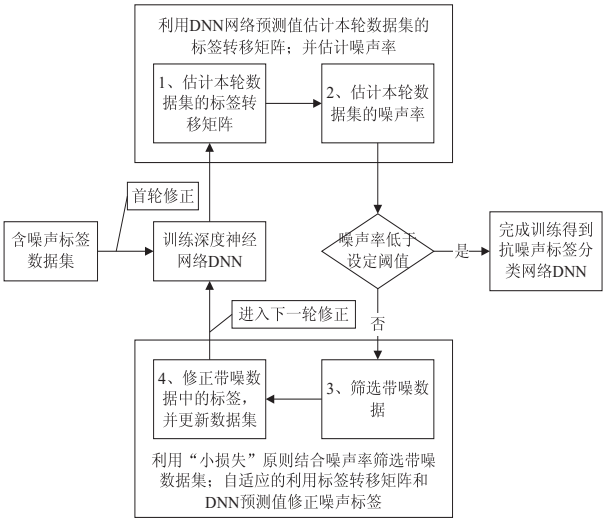

图 1 MCNN 基本结构

下面详细描述了 MCNN 的各个主要组成部分。

2.2 估计标签转移矩阵

噪声数据集的标签转移矩阵代表了数据集中各类

标签之间的错误标注关系[12],现有的方案在估计时往往假设存在一个可信赖标签数据集的子集,与之不同,MCNN 在估计标签转移矩阵时未做假设,而是利用各类标签“锚点样本”的网络预测值来估计标签转移矩阵。用 x^i 表示标签 i 的“锚点样本”,$\bar{\boldsymbol{Y}} = \{\bar{y}_c\}_{c=1}^{C}$ 表示利用带噪数据训练所得网络的输出预测值,其中 $\bar{y}_c = p(\bar{y} = c \mid x)$ 表示样本 x 被预测为 c 类样本的概率,C 表示标签的总类别数目。

标签 i 的“锚点样本” x^i 本质上为该标签的“完美样本”,即神经网络对“锚点样本” x^i 的输出预测值满足如下条件:

$$p(y = i \mid x^i) \to 1, \quad x^i \in \boldsymbol{X}, \boldsymbol{X} \in \boldsymbol{D}$$

$$p(y = j \mid x^i) \to 0, \quad i \neq j$$

其中,$\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ 代表训练数据集,$\boldsymbol{X}$ 代表数据集中的全体样本。因此,该文将数据集所有样本中最大概率标注为某类标签的样本近似看作该类标签的“锚点样本”,即:

$$x^i = \mathrm{argmax}_{x_k} p(\bar{y} = i \mid x_k), x_k \in \boldsymbol{X}$$

在获取到各类标签的“锚点样本”后,标签转移概率可估计如下:

$$p_{ij} = p(\bar{y} = j \mid x^i)$$

相应的标签转移矩阵估计如下:

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & \cdots & p_{1c} \\ \vdots & \ddots & \vdots \\ p_{c1} & \cdots & p_{cc} \end{pmatrix}$$

在 MCNN 的多轮修正过程中,每一轮都估计当前训练数据集 $\boldsymbol{D}_t$ 的标签转移矩阵 $\boldsymbol{P}_t$,其中 t 表示当前训练的轮数。

2.3　计算加权平均噪声率

带数据集的噪声率在诸多任务场景中通常是未知量,然而清除带噪数据需要噪声率作为指标。现有的方案[18]中利用专业人员标注小批量数据并进行专家比对计算噪声率。进而根据小批量数据的噪声率估计整体数据集的噪声率,估计的质量取决于小批量数据的抽样和选择方法,且专业人员的引入增加了任务的额外代价。

该文提出了一种利用标签转移矩阵 $\boldsymbol{P}$ 估计数据集噪声率的方法,无需引入额外的人工标注。具体地:标签转移矩阵的对角线元素 $\{p_{cc}\}_{c=1}^{C}$ 代表了数据集中各类标签被正确标注的概率,利用对角线元素可以实现对噪声率 r 的粗略估计,方法如下:

$$r = \sum_{c=1,2,\cdots,C} (1 - p_{cc})/C$$

考虑到在实际场景中各类型标签的数据量不尽相同,上述结果可能偏离正确值,因此对上述各类型标签的噪声率进行加权求和可提高估计的准确性。对于训练数据集,标签的观察数目 $\boldsymbol{A} = \{a_c\}_{c=1}^{C}$ 以及总的标签数 N 已知。根据标签转移矩阵 $\boldsymbol{P}$ 可近似估计对应无噪声标签数据集中各类标签的数目 $\boldsymbol{B} = \{b_c\}_{c=1}^{C}$,计算方法如下:

$$\boldsymbol{B} = \boldsymbol{A} \times \boldsymbol{P}$$

则估测的无噪声标签数据集中各类标签的权重 $\boldsymbol{W} = \{w_c\}_{c=1}^{C} = \{b_c/N\}_{c=1}^{C}$,数据集的加权平均噪声率 r_w 可求解如下:

$$r_w = \boldsymbol{W} \bullet \boldsymbol{R}$$

其中,$\boldsymbol{R} = \{1 - p_{cc}\}_{c=1}^{C}$ 表示各类标签对应的噪声率向量。

多轮修正过程中,针对第 t 轮的训练数据集 $\boldsymbol{D}_t$,其加权平均噪声率 $r_{w,t}$ 可利用其标签转移矩阵 $\boldsymbol{P}_t$ 求得。

2.4　筛选带噪数据

“小损失”准则表明利用各数据在训练过程中的损失值可以帮助筛选带噪数据。依赖单一网络的数据损失存在“意外记忆”问题。Han 等[16]提出了一种双网络联合学习的方法,然而该方法的缺点在于网络结构比较复杂。

该文依据“小损失”准则实现了一种简单而有效地筛选带噪数据的方法。具体地:通过周期性地改变网络的学习率参数,使得网络循环地由过拟合过渡到欠拟合,由于网络在状态改变时参数随机变化,循环训练过程的单次训练可看作在不同网络上的训练,避免了单一网络因意外记忆而造成的错误筛选问题。假设在循环训练过程中各数据样本的平均损失为 $\boldsymbol{L} = \{l_i\}_{i=1}^{N}$,其中 l_i 表示样本 i 在训练过程中的平均损失,对 $\boldsymbol{L}$ 进行从大到小排序 $\mathrm{Rank}(\boldsymbol{L})$,结合加权平均噪声率 r_w,将序列 $\mathrm{Rank}(\boldsymbol{L})$ 的前 r_w 分位元素所对应的数据筛选出来,筛选出的数据为带噪声标签数据。

在多轮修正的第 t 轮修正中,依据上述方法将本轮的训练数据集 $\boldsymbol{D}_t$ 按照加权平均噪声率 $r_{w,t}$ 划分为带正确标签的数据集 $\boldsymbol{D}_{t,l}$ 和带噪声标签的数据集 $\boldsymbol{D}_{t,n}$,其计算公式如下:

$$\boldsymbol{D}_{t,l} = \boldsymbol{D}_t * (1 - r_{w,t})$$

$$\boldsymbol{D}_{t,n} = \boldsymbol{D}_t * r_{w,t}$$

考虑到筛选存在的误差,在 $\boldsymbol{D}_{t,n}$ 中包含的数据包括真实带噪数据量 $\boldsymbol{S}_t$ 和意外筛选数据量 $\boldsymbol{U}_t$,且有:

$$\boldsymbol{D}_{t,n} = \boldsymbol{S}_t + \boldsymbol{U}_t$$

此时,数据集中实际的带正确标签的数据量可记作 $\boldsymbol{D}_{t,r}$,有:

$$\boldsymbol{D}_{t,r} = \boldsymbol{D}_{t,l} + \boldsymbol{U}_t$$

2.5　联合修正噪声标签

基于清除带噪数据的方法采取将带噪声标签的数

据从训练数据集中清除或给予较小的权重[18],然而该操作有可能意外地去除或忽略一些有用样本[6],从而影响网络的最终性能。为此,结合标签转移矩阵和神经网络的输出预测值,该文提出了一种针对噪声标签的联合修正方法。

在标签转移矩阵中矩阵元素 $\{p_{ij}\}_{i,j=1}^{C}$ 代表标签 i 转换成标签 j 的概率,同时矩阵中的第 j 列向量代表了各类标签可能转换为第 j 类标签的概率,因此对于已知的噪声标签,利用标签转移概率可找到一个最有可能与其对应的正确标签;进一步考虑各类标签在数据集中的占比不同,利用各类标签的权重加以辅助可提高修正的准确性。具体的修正方法如下,设筛选出的噪声标签 $\bar{y}=j$,则标签修正值 $\tilde{y}_1$ 为:

$$\tilde{y}_1 = \arg\max_i(p_{ij} * w_i), \quad i = 1,2,\cdots,C, i \neq j$$

对于利用带噪数据集训练的网络,其网络预测值 $\overline{\boldsymbol{Y}} = \{\bar{y}_c\}_{c=1}^{C} = F(\theta;\ x)$,满足:

在大噪声率下有:

$$p(\arg\max_c(\bar{y}_c) = y) \to 0, \quad c = 1,2,\cdots,C$$

在小噪声率下有:

$$p(\arg\max_c(\bar{y}_c) = y) \to 1, \quad c = 1,2,\cdots,C$$

其中,$\bar{y}_c$ 为样本 x 被判为 c 类样本的概率,y 为其潜在正确标签,$F(\theta)$ 为网络模型表示的映射。为了避免使用训练集中的原始样本,该文对原始样本进行了数据增强。针对不同的数据类型,数据增强的方式可以自由改变。具体表示如下:

$$\{\hat{x}_m\}_{m=1}^{M} = \text{Augment}(x)$$

其中,$\hat{x}_m$ 为样本 x 的增强结果,M 为增强值的数量。该文将 M 个样本增强结果的网络预测值求平均作为噪声标签的另一个修正值 $\tilde{\boldsymbol{Y}}_2$,具体地:

$$\tilde{\boldsymbol{Y}}_2 = \frac{1}{M}\sum_m F(\theta;\ \hat{x}_m)$$

其中,$F(\theta;\ \hat{x}_m)$ 表示增强数据 $\hat{x}_m$ 的网络预测值。

该文利用上述两个修正值对噪声标签进行联合修正,修正结果用 $\tilde{\boldsymbol{Y}}$ 表示:

$$\tilde{\boldsymbol{Y}} = (1 - w(r_w)) * \tilde{\boldsymbol{Y}}_1 + w(r_w) * \tilde{\boldsymbol{Y}}_2$$

其中,$\tilde{\boldsymbol{Y}}_1$ 是利用标签转移矩阵所得到的修正值,$\tilde{\boldsymbol{Y}}_2$ 是根据网络预测所得到的修正值,$w(r_w)$ 是关于噪声率的函数为两个修正值的权重因子,具体地:

$$w(r_w) = 1 - r_w^a$$

其中,a 为超参数,其值的选取依赖于数据集的噪声类型。权重因子满足在数据集的噪声率较大时,使修正值主要依赖于标签转移矩阵修正值;在噪声率较小时,使修正值主要依赖于网络预测修正值。根据数据集的噪声率变化自适应地调整两个修正值的贡献度以提高修正的准确性。

在多轮修正的第 t 轮修正中,利用上述方法对筛选出的带噪声标签数据集 $\boldsymbol{D}_{t,n}$ 进行标签修正,其中对于真实带噪数据量 $\boldsymbol{S}_t$,若假设修正准确率为 λ_t,则该轮正确修正的数据量为 $\lambda_t * \boldsymbol{S}_t$;同理针对意外筛选数据量 $\boldsymbol{U}_t$,若假设意外改错率为 μ_t,则意外纠错数据量为 $\mu_t * \boldsymbol{U}_t$。

因此,由上述可求得第 $t+1$ 轮修正中的训练数据集 $\boldsymbol{D}_{t+1}$ 中实际的带正确标签的数据量,记作 $\boldsymbol{D}_{t+1,r}$,其中:

$$\boldsymbol{D}_{t+1,r} = D_{t,r} + \lambda_t * \boldsymbol{S}_t - \mu_t * \boldsymbol{U}_t$$

且在标签修正过程中正确修正的数据量通常远大于意外纠错数据量,即:$\lambda_t * \boldsymbol{S}_t > \mu_t * \boldsymbol{U}_t$,所以有:

$$\boldsymbol{D}_{t+1,r} > \boldsymbol{D}_{t,r}$$

综上所述,MCNN 的修正算法能够有效地逐轮增加训练数据集中的带正确标签数据量,符合网络设计预期。

2.6 多轮修正的截止条件

该文采取了多轮修正的策略,以尽可能地改善数据集的标签质量水平。多轮修正截止条件的设定可以有多种方法。

实验发现当数据集的噪声率低于 10% 时,继续降低噪声率,训练的网络性能无显著提升。因此,将截止条件设置为:估计所得数据集的噪声率低于 10% 时截止训练。

需要明确指出的是:提出的 MCNN 是一种通用的基于多轮修正使用噪声标签数据进行分类的网络架构,不依赖于所使用的多轮修正截止条件。

3 实 验

3.1 数据集说明

为了验证 MCNN 的性能,在多个数据集上进行了实验,包括 MNIST、FASHION_MNIST 和 CIFAR-10。

3.2 人工加噪

为了获得相应的带噪数据,该文采取了三种广泛使用的加噪方式对数据集进行人工加噪。

(1)对称噪声(Symmetric Noise)[8]:每一类标签以相同概率转换为其他可能的标签。

(2)成对翻转噪声(Pair-flip Noise)[19]:将某一类标签按照某个概率随机替换为其他某一类标签。

(3)模拟真实世界噪声(Simulated Real-words Noise)[20]:考虑数据集的样本特征选取其中最有可能混淆的类别并在它们之间进行替换。

注意在对称噪声(Symmetric Noise)中,每一类标签以相同概率转换为其他可能的标签,转换概率和为噪声率;针对成对翻转噪声(Pair-flip Noise)和模拟真实噪声(Simulated Real-world Noise),如果人工加噪率超过 50% 即噪声标签的数目超过正确标签的数目,

则网络将无法学习到数据集的有效信息[16]。因此,针对不同的噪声类型,该文选取了不同的噪声率进行实验,具体如表 1 所示。

表 1　人工加噪过程中各类型噪声的噪声率

噪声类型	噪声率
Symmetric Noise (S)	50%,70%
Pair-flip Noise (P)	20%,45%
Simulated Real-world Noise (SR)	20%到45%之间的随机值

3.3　对比方案

在对比实验中,将 MCNN 和多种最新的基于噪声标签训练分类神经网络的方案进行了全面深入的对比。对比方案描述如下:

Direct-Training:直接利用带噪数据集训练神经网络,是最基本的对比标准,同时可以直观地看到噪声标签对神经网络性能的影响。

Bootstrapping[9]:使用网络的预测值和原始标签的加权组合作为样本的新标签,然后进行反向传播训练网络。

F-Correction[12]:估计带噪数据集的标签转移矩阵,然后利用该矩阵修改基本分类网络的输出,之后利用修改后结果和噪声标签构成的损失函数训练分类器来达到抗噪声标签的效果。

Co-Teacher[16]:训练两个神经网络。每个网络指导另一个网络在训练中筛除掉可能附有噪声标签的数据,以减小噪声标签对网络参数的影响。

在 Bootstrapping 方案中,其新标签 $\tilde{\boldsymbol{Y}}$ 产生如下:

$$\tilde{\boldsymbol{Y}} = \alpha * \overline{\boldsymbol{Y}} + \beta * \boldsymbol{Y}$$

其中,$\overline{\boldsymbol{Y}}$ 表示网络预测值,$\boldsymbol{Y}$ 表示数据集原始标签,然后利用 $\{\boldsymbol{X},\tilde{\boldsymbol{Y}}\}$ 训练网络。然而其局限性在于,在大噪声率下,$\overline{\boldsymbol{Y}}$ 可信度较低,将其与原始标签相组合不仅无法改善标签质量,甚至还可能污染原本正确的标签,同时其权值参数 α,β 根据经验设定,对标签质量的改善能力有限。在所提出的标签修正算法中,修正结果 $\tilde{\boldsymbol{Y}}$ 产生如下:

$$\tilde{\boldsymbol{Y}} = (1 - w(r_w)) * \tilde{\boldsymbol{Y}}_1 + w(r_w) * \tilde{\boldsymbol{Y}}_2$$

其中,$\tilde{\boldsymbol{Y}}_1$ 为利用标签转移矩阵获得的修正标签,$\tilde{\boldsymbol{Y}}_2$ 为经过优化后的网络预测值,因此,当 $\tilde{\boldsymbol{Y}}_2$ 可信度降低时,仍可以利用 $\tilde{\boldsymbol{Y}}_1$ 对噪声标签进行修正,同时利用 $w(r_w)$ 动态调节两者的权值提高了修正的准确性。另外,该文只针对筛选出的错误标签进行处理,有效避免了污染正确标签的问题。

在 F-Correction 中,该文利用估计所得的标签转移矩阵 $\boldsymbol{P}$ 修改分类网络的输出预测值,具体地:

$$\tilde{\boldsymbol{Y}} = P \times \overline{\boldsymbol{Y}}$$

之后利用 $\tilde{\boldsymbol{Y}}$ 和数据集原始标签 $\boldsymbol{Y}$ 构成的损失函数训练分类器。然而标签转移矩阵 $\boldsymbol{P}$ 存在估计误差,上述利用 $\boldsymbol{P}$ 做矩阵乘法所得到的 $\tilde{\boldsymbol{Y}}$ 很容易因误差而偏离预期结果。该文利用标签转移矩阵 $\boldsymbol{P}$ 修正噪声标签具体有:

$$\tilde{y}_1 = \arg\max_i (p_{ij} * w_i), \quad i = 1,2,\cdots,C, i \neq j$$

其只利用矩阵元素的相对大小而不做矩阵乘法,误差容限较大,所以具有一定的优越性。

在 Co-Teacher 方案中,将原始训练数据集 $\boldsymbol{D}_t$ 中的带噪声标签数据 $\boldsymbol{D}_{t,n}$ 依据“小损失”准则筛选出来并去除,并利用剩余的带正确标签的数据集 $\boldsymbol{D}_{t,l}$ 训练网络,其缺点在于将 $\boldsymbol{D}_{t,n}$ 去除的做法严重削减了样本数量,在大噪声率条件下表现较差。该文提出的网络框架则是将带噪声标签数据 $\boldsymbol{D}_{t,n}$ 逐轮修正,有效地扩大了带正确标签的数据集 $\boldsymbol{D}_{t,l}$,因此在大噪声率条件下具有一定的优势。

综上所述,提出的 MCNN 网络框架较现有的方法具有一定的优势。

3.4　实验设置

(1)针对不同的数据集,选取了不同神经网络模型进行训练,具体信息如表 2 所示。

表 2　数据集训练所用分类网络

数据集	网络架构
MNIST	Le-Net[12]
FASHION-MNIST	ResNet14[19]
CIFAR-10	ResNet18[20]

(2)关于筛选带噪数据步骤中周期性改变网络的学习率参数的相关设置:循环训练可解决网络的“意外记忆”问题,然而循环次数过多会增加训练时长,平衡之下选取循环次数为 3。

(3)关于修正策略的权重因子中参数的选取,其依赖于数据集的噪声类型。鉴于在相同噪声率条件下,在对称噪声的数据集上训练的网络具有较好的分类性能,而利用带有成对翻转噪声和模拟真实噪声的数据集训练的网络分类性能则较差,因此在对抗对称噪声时 a 选取 2;在成对翻转噪声和模拟真实噪声时 a 选取 0.7。

3.5　性能比较指标

从分类精度上将提出的方案和各种对比方案进行了比较和评估。模型在测试集上的分类精度定义为:

$$\frac{\text{预测正确的样本数量}}{\text{输入测试的样本数量}}$$

表 3 给出了各种方案在 MNIST、FASHION-MNIST 和 CIFAR-10 的测试集上的分类精度比较,其中 S 代表对称噪声,P 代表成对翻转噪声,S-R 代表模

拟真实噪声,相应的数字代表人工加噪的噪声率。

表 3 各种方案在 MNIST、FASHION-MNIST 和 CIFAR-10 的测试集上的分类精度比较 %

方案		S-0.5	S-0.7	P-0.2	P-0.45	S-R
MNIST	Direct-training	92.25	76.10	80.20	60.32	76.03
	Bootstrapping	94.90	76.55	85.37	62.12	80.44
	F-Correction	97.30	86.21	88.75	80.24	85.48
	Co-Teacher	96.25	93.32	91.60	87.90	88.20
	MCNN	98.81	96.70	94.90	93.30	93.26
FASHION-MNIST	Direct-training	90.62	75.71	78.18	57.29	75.18
	Bootstrapping	92.91	76.80	80.50	60.70	79.40
	F-Correction	95.62	85.10	87.04	81.30	85.60
	Co-Teacher	94.10	90.53	88.20	86.10	86.80
	MCNN	97.91	96.35	91.24	89.73	90.24
CIFAR-10	Direct-training	76.21	59.72	61.18	49.50	58.18
	Bootstrapping	83.11	61.61	73.60	49.05	68.31
	F-Correction	84.66	77.37	78.32	60.16	75.01
	Co-Teacher	82.32	80.53	79.48	77.62	79.32
	MCNN	88.42	87.80	82.24	80.73	82.79

3.6 实验结果与分析

从表 3 中可以得出如下结论:

(1)MCNN 在多种不同的噪声类型和 DNN 网络结构下都取得了最好的分类性能,充分说明了该方案的优越性和健壮性。

(2)由表 3 可以看出,Bootstrapping[9] 在小噪声率情况下如 S-0.5 和 P-0.2 表现良好,而在大噪声率情况如 S-0.7 和 P-0.45 则表现较差,原因在于在大噪声率下,网络的预测值可信度较低,Bootstrapping 将其与噪声标签相组合不仅无法改善标签质量,甚至还可能污染原本正确的标签。提出的 MCNN 只对筛选出的错误标签进行处理,有效避免了污染正确标签的问题,因此最终的模型性能优于前者,针对各类数据集在 S-0.7 和 P-0.45 上,MCNN 最终的模型在分类精度上有超过 20% 的提升。

(3)相比于 F-Correction[12],前者利用标签转移矩阵与数据样本的网络预测值做矩阵乘积,然而标签转移矩阵在估计时通常存在误差,在数据集数据量较大时,存在误差积累。该文提出的 MCNN 只单次利用标签转移矩阵推测加权平均噪声率误差累积较小;同时,在标签修正阶段,该文只利用矩阵元素的相对大小来提供一个噪声标签修正值而不做矩阵乘法,误差容限较大。因此,最终训练的模型在各类型噪声下分类精度都有一定的提升。

(4)Co-Teacher[16] 通过清洁数据集来对抗噪声标签,在小噪声率条件下,Co-Teacher 和 MCNN 大体上性能相同,都实现了较好的抗噪声性能,例如在 MNIST 和 FASHION-MNIST 上针对 S-0.5 都达到了超过 90% 的分类精度。而在大噪声率条件下,MCNN 的性能优于 Co-Teacher,原因在于 Co-Teacher 采用将带噪数据清除,在大噪声率条件下严重削减了样本数量,而 MCNN 则采取将噪声数据修正减少了数据量的损失,因此在大噪声率下模型的性能优于前者,例如在 CIFAR-10 数据集上针对 S-0.7 在分类精度上有接近 7% 的提升。

3.7 MCNN 的可行性分析

提出的 MCNN 能够改善数据集质量的因素主要在于两点:错误标签筛选和多轮修正。关键之处在于以下两个方面:

(1)MCNN 能够实现对数据集标签转移矩阵的准确估计,同时可以精确地估计数据集的噪声率用以为筛选带噪数据提供基准。

(2)MCNN 能够实现对带噪数据的准确筛选即具有较高的带噪数据筛选精度,并且提出的修正方法可以实现对标签的“正向修正”。

3.7.1 仿真结果

选取 MCNN 在 MNIST 数据上针对模拟真实噪声(S-R)的实验结果详细阐述 MCNN 的可行性。为此,采用如下 5 个 MCNN 运行过程中的中间性能量度来定量展示 MCNN 提升数据集质量的两个关键方面。

(1)标签转移矩阵估计误差。定义如下:

$$\frac{\| \text{标签转移矩阵实际值} - \text{标签转移矩阵估计值} \|_1}{\| \text{标签转移矩阵实际值} \|_1}$$

(2)噪声率估计误差。定义如下:

$$\| \text{噪声率实际值} - \text{加权平均噪声率} \|_1$$

(3)带噪数据筛选精度。定义如下:

$$\frac{\text{筛选出的带有噪声标签的样本数目}}{\text{筛选出的总样本数目}}$$

(4)修正准确率。定义如下:

$$\frac{\text{将筛选出的噪声标签正确修正的绝对数量}}{\text{筛选出的总数据数目}}$$

(5)意外改错率。定义如下:

$$\frac{\text{将意外筛选的正确标签错误修改的绝对数目}}{\text{筛选出的总数据数目}}$$

图 2 和图 3 给出了在 MNIST 数据集上,MCNN 在对抗模拟真实噪声(S-R)时的实验情况,经过两轮修正(第三轮截止训练)数据集噪声率下降至设定阈值。

3.7.2　标签转移矩阵和噪声率的估计情况

图 2 中结果表明,MCNN 能够实现对数据集标签转移矩阵的准确估计,在两轮修正过程中标签转移矩阵的估计误差波动在 2.24% ~ 3.32% 之间。同样地,MCNN 提出的加权平均噪声率可以较准确地估计数据集的噪声率,由图 2 可以看出在 3 轮训练过程中估计误差波动在 1.20% ~ 1.91% 之间。

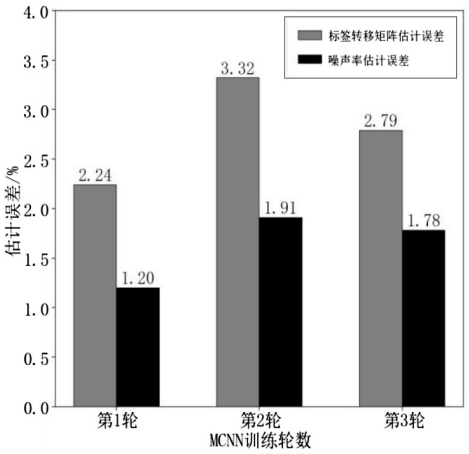

图 2　MCNN 对抗 S-R 时标签转移矩阵和噪声率的估计误差

3.7.3　筛选带噪数据和修正情况

MCNN 在处理模拟真实噪声(S-R)时的筛选带噪数据和标签修正情况如图 3 所示。由图中数据可得:在筛选带噪数据上 MCNN 表现良好,带噪数据筛选精度波动在 81.64% 左右;在噪声标签修正方面,可以看出 MCNN 虽无法实现对筛选出的全部的噪声标签进行正确修正,但是在两次修正过程中修正准确率均远大于意外改错率,即 MCNN 在各轮训练上均可实现对标签的“正向修正”。

综上所述,提出的网络框架 MCNN 能够实现数据集质量提升的两个关键方面,因此在经过多轮修正后数据集质量水平可达到预期目标,训练所得分类网络能够实现较理想的分类性能。

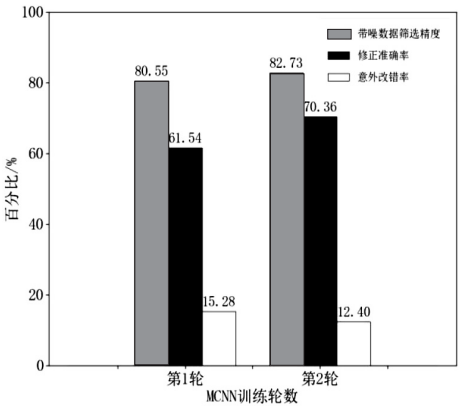

图 3　MCNN 对抗 S-R 时的带噪数据筛选精度、修正准确率和意外改错率

4　结束语

提出了一种通用的基于多轮修正的抗噪声标签神经网络框架 MCNN,以解决分类任务中噪声标签对网络性能的影响。MCNN 首先基于神经网络的学习特性将带噪声标签数据按噪声率比例筛选出来,之后利用修正方案修正噪声标签逐轮提高数据集质量,最后利用高质量的数据集训练网络。在多个真实数据集和多种噪声类型下与其他基准方案进行了比较,结果表明提出的 MCNN 有较好的抗噪声标签能力。利用多轮修正的方法对数据集中的噪声标签进行尽可能的修正,然而对于筛选出的带噪数据可能不必要全部进行修正,转而利用选择策略[21]选取其中对网络参数有显著影响的样本进行修正,则可优化多轮修正的轮数和代价。之后的工作将是发掘、设计一些可行的选择策略来选取数据,这是文中工作的一个后续展望。

参考文献:

[1] CORDEIRO F R, CARNEIRO G. A survey on deep learning with noisy labels: how to train your model when you cannot trust on the annotations[C]//Proceedings of 33rd conference on graphics, patterns and images. Pernambuco: IEEE, 2020: 9-16.

[2] ARPIT D, JASTRZEBSKI S, BALLAS N, et al. A closer look at memorization in deep networks[C]//Proceedings of the international conference on machine learning. New York: PMLR, 2017: 233-242.

[3] STEINHARDT J, KOH P W, LIANG P S. Certified defenses for data poisoning attacks[J]. Advances in Neural Information Processing Systems, 2017, 30: 3518-3530.

[4] CHEN P, LIAO B B, CHEN G, et al. Understanding and utilizing deep neural networks trained with noisy labels[C]// Proceedings of the international conference on machine learning. New York: PMLR, 2019: 1062-1070.

[5] CHEN X L, SHRIVASTAVA A, GUPTA A. NEIL: extracting visual knowledge from web data[C]//Proceedings of the international conference on computer vision. Sydney: IEEE, 2013: 1409-1416.

[6] 宫 辰,张 闯,王启舟. 标签噪声鲁棒学习算法研究综述[J]. 航空兵器,2020,27(3):20-26.

[7] KRAUSE J, SAPP B, HOWARD A, et al. The unreasonable effectiveness of noisy data for fine-grained recognition[C]//Proceedings of the 14th European conference on computer vision. Amsterdam: Springer, Cham, 2016: 301-320.

[8] HUANG J, QU L, JIA R, et al. O2u-net: a simple noisy label detection approach for deep neural networks[C]//Proceedings of the international conference on computer vision. Seoul: IEEE, 2019: 3326-3334.

[9] SCOTT R, HONGLAK L, DRAGOMIR A, et al. Training deep neural networks on noisy labels with bootstrapping[C]//Proceedings of the international conference on learning representations. San Diego: IEEE, 2015

[10] 余孟池,牟甲鹏,蔡 剑,等. 噪声标签重标注方法[J]. 计算机科学,2020,47(6):79-84.

[11] JINDAL I, NOKLEBY M, CHEN X. Learning deep networks from noisy labels with dropout regularization[C]//Proceedings of the 16th international conference on data mining. Barcelona: IEEE, 2016: 967-972.

[12] PATRINI G, ROZZA A, KRISHNA M A, et al. Making deep neural networks robust to label noise: a loss correction approach[C]//Proceedings of the conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 1944-1952.

[13] HENDRYCKS D, MAZEIKA M, WILSON D, et al. Using trusted data to train deep networks on labels corrupted by severe noise[J]. Advances in Neural Information Processing Systems, 2018, 30: 10456-10465.

[14] HAN B, YAO J, NIU G, et al. Masking: a new perspective of noisy supervision[J]. Advances in Neural Information Processing Systems, 2018, 31: 5836-5846.

[15] ERAN M, SHAI S S. Decoupling "when to update" from "how to update"[J]. Advances in Neural Information Processing Systems, 2017, 124: 960-970.

[16] HAN B, YAO Q, YU X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[J]. Advances in Neural Information Processing Systems, 2018, 31: 8527-8537.

[17] YU X, HAN B, YAO J, et al. How does disagreement help generalization against label corruption[C]//Proceedings of the international conference on machine learning. Long Beach: PMLR, 2019: 7164-7173.

[18] REN M, ZENG W, YANG B, et al. Learning to reweight examples for robust deep learning[C]//Proceedings of the international conference on machine learning. Stockholm: PMLR, 2018: 4334-4343.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 770-778.

[20] ZHANG C, BENGIO S, HARDT M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107-115.

[21] LIU P, WANG L, RANJAN R, et al. A survey on active deep learning: from model driven to data driven[J]. ACM Computing Surveys, 2022, 54(10s): 1-34.