计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600041

噪声标签重标注方法

余孟池　牟甲鹏　蔡　剑　徐　建
南京理工大学计算机科学与工程学院　南京 210094
(2246782556@qq.com)

摘　要　样本标签的完整性对于有监督学习问题的分类精度有着显著影响,然而在现实数据中,由于标注过程的随机性和标注人员的不专业性等因素,数据标签不可避免地会受到噪声污染,即样本的观测标签不同于真实标签。为降低噪声标签对分类器分类精度的负面影响,文中提出一种噪声标签纠正方法,该方法利用基分类器对观测样本进行分类并估计噪声率,以识别噪声标签数据,再利用基分类器的分类结果对噪声标签样本进行重新标注,得到噪声标签样本被修正后的样本数据集。在合成数据集与真实数据集上的实验结果表明,该重标注算法在不同基分类器和不同噪声率干扰下对分类结果都有一定的提升作用,在合成数据集上对比无降噪声算法,其正确率提升5%左右,而在CIFAR和MNIST数据集上的高噪声率环境下,该重标注算法的$F1$值比Elk08和Nat13平均高7%以上,比无噪声算法高53%。

关键词:噪声标签学习;重标注标签;逻辑回归;朴素贝叶斯
中图法分类号　TP301

Noisy Label Classification Learning Based on Relabeling Method

YU Meng-chi, MU Jia-peng, CAI Jian and XU Jian
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract　The integrity of sample labels has a significant impact on the accuracy of supervised learning algorithms. However, in real data, due to the unprofessional and random nature of the labeling process, the label of the dataset is inevitably polluted by noise, i.e. the assigned label of sample is different from its real label. In order to reduce the negative impact of noise labels on the classification accuracy of classifiers, this paper proposes a noise label correction approach. It firstly identifies the noise label data by applying the base classifier to classify the samples and estimating the noise rate to identify noisy label data, and then uses the base classifier to relabel the noisy samples. As a result, the noisy samples are relabeled to obtain a sample dataset in which the noisy samples are corrected. Experiments on synthetic datasets and real datasets show that the relabel algorithm has a certain improvement effect on classification results under different base classifiers and different types of noise rate interference. Compared with the base classifier, the accuracy of relabel algorithm is improved by about 5% in the synthetic dataset, while in the high noise environment of CIFAR and MNIST datasets, the $F1$ score of the proposed algorithm is 7% higher than that of Elk08 and Nat13 on average, and is improved by 53% compared with base classifier.

Keywords　Noisy label learning, Relabeling label, Logistic Regression, Naive Bayes

1　引言

传统的监督学习分类问题通常假设数据集的标签是完整的,即每个数据集样本都存在无噪声的正确标签。然而在现实世界中,由于标签标注过程的随机性,样本标签容易受噪声污染而不准确。噪声数据的产生通常与数据集的获取途径有关。例如,在对原始数据标注的过程中,提供给标注人员的样本数据信息量不足,从而导致标注人员将样本错误分类;又或者由于分类过程本身就是一个主观过程或是标注人员专业知识有限,无法保证分类的正确性。目前,各种流行的数据标注平台也是噪声数据的来源之一,例如Amazon的Amazon Mechanical Turk、数据堂、京东微工等数据服务平台,这些标注平台利用广大注册用户实现众包式的数据标注工作。由于标注者的专业性限制或个人差异,通过这种途径得到的数据标签并不完全符合真实情况,而且不同标注者对同一样本的看法可能不同,从而导致同种样本有不同的标签结果。数据集中的噪声可以根据噪声产生的位置分为特征噪声和标签噪声,一般标签中的噪声比特征中的噪声对模型性能的影响更大[1]。在二元分类中,根据正例数据集和负例数据集中噪声分布的特征,Kheta[2]提出了PU(Positive-unlabeled)学习问

到稿日期:2019-06-11　返修日期:2019-09-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61872186,61802205,91846104)
The work was supported by the National Natural Science Foundation of China (61872186,61802205,91846104).
通信作者:徐建(dolphin.xu@njust.edu.cn)

题。PU学习是数据集中只有一部分正例训练样本有标签而其他样本都不带标签的一种二元分类任务。针对PU学习问题,可以将所有未标注样本当作负例样本,这样PU学习问题就转化为带噪声的二元分类问题。噪声标签数据不仅会对分类器模型的分类准确性产生严重的负面影响,同时也会增加分类器的复杂度。因此,设计适应噪声标签数据的分类学习算法具有重要的研究意义和应用价值。

对于含有噪声标签的分类问题,Frenay等[3]归纳了多种解决策略,包括噪声清理算法[4-7]、噪声标签鲁棒方法[8-9]和噪声标签模型化方法[10]。噪声标签鲁棒方法依赖于对标签噪声具有鲁棒能力的算法。标签噪声对该类分类器的学习过程不会造成太大的影响,例如在二元分类的经验风险最小化问题中,使用损失函数衡量错误分类的损失,通过最小化样本的最小损失学习分类器,常见的损失有0-1损失。对于均匀标签噪声,0-1损失和最小平方损失是抗噪声标签的,而其他的损失函数即使在噪声分布均匀情况下也不是抗噪声标签的,如指数损失、对数损失和hinge损失。机器学习中的大多数学习算法都不完全是抗噪声标签的,并且只在训练数据被少量标签噪声干扰时很有效。随着深度学习的发展,在图像分类问题中常使用神经网络来解决噪声标签图像问题[11-13]。例如Mnih等[14]提出将噪声模型并入神经网络,但其仅考虑了二元分类,并且假定噪声属于对称标签噪声(即噪声标签的分布独立于真实标签)。

使用噪声清理策略解决噪声标签学习问题通常需要两步:1)估计噪声率 ρ_1 和 ρ_0;2)使用噪声率 ρ_1 和 ρ_0 预测。为估计噪声率,Scott[15]提出一种效率较高的噪声率估计方法,但其估计性能表现较差。Liu等[16]通过重要性权值重写修改损失函数,但重写的权值来源于预测概率,因此可能会对不准确的估计比较敏感。Natarajan等[17]没有提出估计噪声的方法,而是将噪声率视为交叉验证过程中优化的参数。Natarajan提出两种方法来修改损失函数,第一种方法是从噪声分布中构建正确分布的无偏估计器,但该估计器即使在原有损失函数是凸函数的情况下仍有可能是非凸函数;第二种方法是建立标签依赖的损失函数,使得对于0-1损失,Nat13算法的最小风险与基准数据分布的风险相等。Northcutt等[18]提出从信任的样本中学习(Learning with confide-nt examples)的概念,按照基分类器对噪声数据的分类概率计算 ρ_0,ρ_1,π_0,π_1 等变量,根据基分类器对每个样本的预测结果识别噪声标签数据,并删除该部分样本,该过程被称为按秩剪枝。

在基分类器性能较差时,Northcutt判定噪声样本的过程中会将过多样本划分为噪声样本,因此该剪枝过程会丢失大量数据,导致分类性能较差。为解决样本量不足的问题,本文沿用Northcutt判断噪声标签的思路,为噪声标签数据重新标注正确标签以充分利用所有样本信息,并分析多类分类情况下的重标注过程。

2 噪声标签问题描述

本文根据基分类器在噪声数据下的分类结果分析数据分布情况,从而识别噪声标签样本,并完成样本标签的重标注。给定 n 个观测训练样本 $x\in R^D$,每个样本带有被噪声干扰后的观测标签 $s\in\{0,1\}$ 和隐藏的真实标签 $y\in\{0,1\}$,需要找到二元分类器 f 满足映射 $x\to y$。然而,如果使用观测数据 (x,s) 训练分类器,将会得到映射 $x\to s$,从而得到 $g(x)=P(\hat{s}=1|x)$。在二元分类中,定义观测噪声正例数据集合和噪声负例数据集合分别为 $\tilde{P}=\{x|s=1\}$ 和 $\tilde{N}=\{x|s=0\}$。隐藏的正确正例集合和正确负例集合分别为 $P=\{x|y=1\}$ 和 $N=\{x|y=0\}$。用"1"表示正例集,"0"表示负例集。定义隐藏的真实数据为 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$。假设一个类条件分类噪声过程(Classification Noise Process,CNP)(也被称为随机噪声分类(Random Classification Noise,RCN)模型)将正确标签 y 映射到观测标签 s,使得 P 中的每个标签以概率 ρ_1 被翻转,而 N 中每个标签以概率 ρ_0 被翻转,最后得到含噪声标签的被干扰数据集为 $D\rho=\{(x_1,s_1),(x_2,s_2),\cdots,(x_n,s_n)\}$。在噪声标签学习中,通常使用表1中定义的变量。

表1 二元分类中各种噪声率变量表示

Table 1 Different noise rate variables in binary classification

变量表示	含义
$\rho_0=P(s=1\|y=0)$	负例样本被误标注为正例标签的概率
$\rho_1=P(s=0\|y=1)$	正例样本被误标注为负例标签的概率
$\pi_0=P(y=1\|s=0)$	观测负例样本集中误标注样本的概率
$\pi_1=P(y=0\|s=1)$	观测正例样本集中误标注样本的概率

噪声率 $\rho_1=P(s=0|y=1)$ 表示 P 中样本被误标注为负例的比例,噪声率 $\rho_0=P(s=1|y=0)$ 表示 N 中样本被误标注为正例的比例($\rho_1+\rho_0<1$,否则误标注样本将比正确标注样本更多)。此外,如果用 $p_{s1}=P(s=1)$ 表示观测样本中正例标签的比例,用 $p_{y1}=P(y=1)$ 表示正确样本中正例标签的比例,则根据贝叶斯定理,反转噪声率可通过式(1)计算:

$$\pi_1=P(y=0|s=1)=\frac{\rho_0(1-p_{y1})}{p_{s1}}$$
$$\pi_0=P(y=1|s=0)=\frac{\rho_1 p_{y1}}{(1-p_{s1})} \tag{1}$$

3 重标注算法

3.1 噪声标签样本识别

噪声率 $\rho_1=P(s=0|y=1)$ 表示真实标签为1的样本被误标记为0的概率,即正确标签为1的样本集中其观测标签为0的样本数量比例。用以下变量表示各种情况下样本的数量:$\tilde{P}_{y=1}$ 表示观测标签为1,真实标签为1的样本;$\tilde{N}_{y=1}$ 表示观测标签为0,真实标签为1的样本;$\tilde{P}_{y=0}$ 表示观测标签为1,真实标签为0的样本;$\tilde{N}_{y=0}$ 表示观测标签为0,真实标签为1的样本。

因为样本的真实分布未知,所以使用基分类器的分类结果 $g(x)=P(\hat{s}=1|x)$ 判断样本的真实标签。考虑使用下界阈值 $LB_{y=1}$ 判断样本的真实标签是否为1。当观测样本在基分类器 $g(x)$ 上的预测结果大于该下界阈值时,可以假设该观测样本的真实标签为1。同样,使用上界阈值 $UB_{y=0}$ 判断观测样本真实标签是否为0。因此,以上各区间内的样本数量可以通过以下方式计算:

$$\begin{aligned}\widetilde{P}_{y=1}&=\{x\in\widetilde{P}\mid g(x)\geqslant LB_{y=1}\}\\ \widetilde{N}_{y=1}&=\{x\in\widetilde{N}\mid g(x)\geqslant LB_{y=1}\}\\ \widetilde{P}_{y=0}&=\{x\in\widetilde{P}\mid g(x)\leqslant UB_{y=0}\}\\ \widetilde{N}_{y=0}&=\{x\in\widetilde{N}\mid g(x)\leqslant UB_{y=0}\}\end{aligned}\tag{2}$$

其中,阈值被设定为正、负例样本在基分类器上的分类概率 $g(x)=P(s=1\mid x)$ 的期望值:

$$\begin{aligned}LB_{y=1}&:=P(\hat{s}=1\mid s=1)=E_{x\in\widetilde{P}}[g(x)]\\ UB_{y=0}&:=P(\hat{s}=1\mid s=0)=E_{x\in\widetilde{N}}[g(x)]\end{aligned}\tag{3}$$

因此,噪声率的估计值 $\hat{\rho}_1$ 和 $\hat{\rho}_0$ 的计算过程如下:

$$\begin{aligned}\hat{\rho}_1&:=\frac{|\widetilde{N}_{y=1}|}{|\widetilde{N}_{y=1}|+|\widetilde{P}_{y=1}|}\\ \hat{\rho}_0&:=\frac{|\widetilde{P}_{y=0}|}{|\widetilde{P}_{y=0}|+|\widetilde{N}_{y=0}|}\end{aligned}\tag{4}$$

由贝叶斯定理,根据噪声率的估计值可以推导出反转噪声率:

$$\begin{aligned}\hat{\pi}_1&=\frac{\hat{\rho}_0}{p_{s1}}\frac{1-p_{s1}-\hat{\rho}_1}{1-\hat{\rho}_1-\hat{\rho}_0}\\ \hat{\pi}_0&=\frac{\hat{\rho}_1}{1-p_{s1}}\frac{p_{s1}-\hat{\rho}_0}{1-\hat{\rho}_1-\hat{\rho}_0}\end{aligned}\tag{5}$$

其中,$p_{s1}=P(s=1)$ 表示观测样本集中正例样本的个数。由于反转噪声率表示观测正、负例样本中真实标签为 0 或 1 的概率,因此 $\hat{\pi}_1|\widetilde{P}|$ 表示观测正例样本集中真实标签为 0 的样本数,即观测正例样本集中的噪声样本数。同理,$\hat{\pi}_0|\widetilde{N}|$ 表示观测负例样本集中真实标签为 1 的样本数,即观测负例样本集中的噪声样本数。最后,根据每个样本基分类器 $g(x)$ 的预测值,将样本升序排列,在观测正例样本集 $\widetilde{P}$ 中,前 $\hat{\pi}_1|\widetilde{P}|$ 个样本被视为正例样本集中的噪声标签样本;在观测负例样本集 $\widetilde{N}$ 中,后 $\hat{\pi}_0|\widetilde{N}|$ 个样本被视为负例样本集中的噪声标签样本。

3.2 噪声标签样本的重标注过程

在识别出噪声标签样本后,根据每个样本在基分类器 $g(x)=P(s=1\mid x)$ 中的预测概率值,将样本升序排序。在观测正例样本集 $\widetilde{P}$ 中,将前 $\hat{\pi}_1|\widetilde{P}|$ 个样本的标签重标注为 0;在观测负例样本集 $\widetilde{N}$ 中,将后 $\hat{\pi}_0|\widetilde{N}|$ 个样本标签重标注为 1。重新标注后的正例样本集 $\widetilde{P}_{\text{relabel}}$ 和负例样本集 $\widetilde{N}_{\text{relabel}}$ 分别表示为:$\widetilde{P}_{\text{relabel}}=\{x\in\widetilde{P}\mid g(x)\geqslant g_{\hat{\pi}_1|\widetilde{P}|}\}\cup\{x\in\widetilde{N}\mid g(x)\geqslant g_{\hat{\pi}_0|\widetilde{N}|}\}$ 和 $\widetilde{N}_{\text{relabel}}=\{x\in\widetilde{N}\mid g(x)\leqslant g_{\hat{\pi}_0|\widetilde{N}|}\}\cup\{x\in\widetilde{P}\mid g(x)\leqslant g_{\hat{\pi}_1|\widetilde{P}|}\}$,其中 $g_{\hat{\pi}_1|\widetilde{P}|}$ 表示观测正例样本集中第 $\hat{\pi}_1|\widetilde{P}|$ 小的 $g(x)$ 值,$g_{\hat{\pi}_0|\widetilde{N}|}$ 表示观测负例样本集中第 $\hat{\pi}_0|\widetilde{N}|$ 大的 $g(x)$ 值。

在多类分类情况下,样本标签总种类数不止两种,此时对噪声样本的标签重新标记需要考虑样本最可能属于哪类标签并分配该标签。噪声样本重标注的标签需要根据基分类器对所有样本的分类结果选择。因此,基分类器在预测所有样本数据时需要记录样本属于每个类别的概率,最终得到一个分类结果矩阵 $\boldsymbol{psx}=\{p_{ij}\mid i\in N,j\in K\}$,$\boldsymbol{psx}$ 是一个 $|N|\times|K|$ 的概率矩阵($|N|$ 为样本数,$|K|$ 为标签种类数),其中的元素值表示基分类器对样本在不同类别上的分类概率,矩阵第 i 行 $p_i=(p_{i1},p_{i2},\cdots,p_{ik})$ 表示样本 x_i 在基分类器 $g(x)$ 下属于各类标签的概率,p_{ij} 表示样本 x_i 属于 k_j 类的概率。当样本 x 被判定为噪声标签后,利用概率矩阵 $\boldsymbol{psx}$ 为样本 x 重新分配标签。重标注算法认为基分类器预测概率最大的标签应该为该样本的正确标签,因此将该样本重标注为该标签。即对于噪声标签样本 x_i,其重标注的标签为 $y_i^{\text{relabel}}=k_{\max}$($k_{\max}=\text{argmax } psx_i$),其中 $k_{\max}$ 为样本 x_i 在基分类器分类概率中除该样本原有噪声标签 s_i 外概率最大值所属的标签类别。

重标注后得到的数据集即为修正噪声标签后的正确数据集。在重标注后的数据集上重新训练分类器。在总样本数为 n 的数据集中,重标注算法在计算 $g_{\hat{\pi}_1|\widetilde{P}|}$ 和 $g_{\hat{\pi}_0|\widetilde{N}|}$ 时判断每个样本的分类概率值,因此重标注算法识别噪声样本的时间复杂度为 $O(n)$。假设基分类器的时间复杂度为 $O(T)$,则整个分类过程的总时间复杂度为 $O(n)+O(T)$。完整的算法流程如算法 1 所示。

算法 1 噪声标签重标注算法

输入:样本 X,噪声标签 s,基分类器 clf

输出:输出重标注后的数据集(X,y^{relabel})

1. 基分类器 clf 对样本预测 clf.fit(X,s),得到样本预测概率 g(x)=P(s=1|x)。
2. 分别取正例集合和负例集合所有样本的预测概率期望值作为下阈值和上阈值 $LB_{y=1}=E_{x\in\widetilde{P}}[g(x)]$,$UB_{y=0}=E_{x\in\widetilde{N}}[g(x)]$。
3. 根据式(2)得到 $\widetilde{P}_{y=1}$,$\widetilde{N}_{y=1}$,$\widetilde{P}_{y=0}$,$\widetilde{N}_{y=0}$。
4. 根据式(4)得到噪声率 $\hat{\rho}_1$,$\hat{\rho}_0$。
5. 根据式(5)得到 $\hat{\pi}_1$,$\hat{\pi}_0$。
6. $\widetilde{P}$ 中,前 $\hat{\pi}_1|\widetilde{P}|$ 个样本即为正例样本集中的噪声标签样本,在 $\widetilde{N}$ 中,后 $\hat{\pi}_0|\widetilde{N}|$ 个样本即为负例样本集中的噪声标签样本

//重标注过程

7. 对于二元分类:重标注后的正例集为 $\widetilde{P}_{\text{relabel}}=\{x\in\widetilde{P}\mid g(x)\geqslant g_{\hat{\pi}_1|\widetilde{P}|}\}\cup\{x\in\widetilde{N}\mid g(x)\geqslant g_{\hat{\pi}_0|\widetilde{N}|}\}$,重标注后的负例样本集为 $\widetilde{N}_{\text{relabel}}=\{x\in\widetilde{N}\mid g(x)\leqslant g_{\hat{\pi}_0|\widetilde{N}|}\}\cup\{x\in\widetilde{P}\mid g(x)\leqslant g_{\hat{\pi}_1|\widetilde{P}|}\}$。
8. 对于多类分类:噪声标签样本 x_i 的重标注标签为 $y_i^{\text{relabel}}=k_{\max}$($k_{\max}=\text{argmax } psx_i$)。

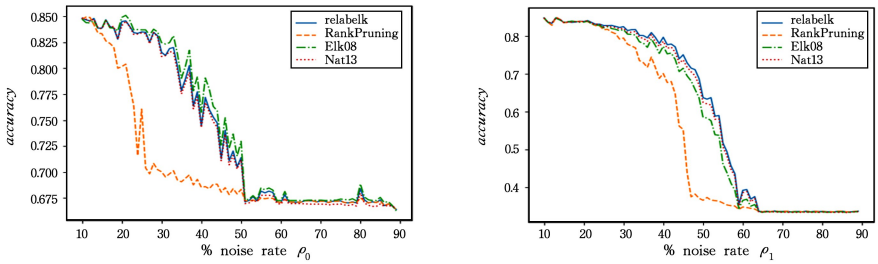
4 实验

本节通过不同场景下的实验验证本文提出的噪声标签重标注方法的有效性。该方法适用于不同的基分类器,实验选取了两种基分类器算法,分别为逻辑回归(Logistic Regression)和朴素贝叶斯(Naive Bayes)算法,所有噪声标签处理算法分别使用这两种基分类器运行。实验选取了 Northcutt[18] 的 RankPruning、Elkan 等[19] 的 Elk08 和 Natarajan[17] 的 Nat13 算法作为对比算法。在合成数据集中使用正确率(accuracy)作为评价指标,在真实数据集中使用 F1-score 值作为评价指标。

4.1 合成数据集

在合成数据集实验中，本文利用 Python 的 Numpy 库生成多元高斯分布数据，正例数据集和负例数据集通过向多元高斯分布指定不同的均值(mean)和协方差(cov)参数生成。正例集 P 和负例集 N 分别服从不同的多元高斯分布：$P\sim N(\boldsymbol{\mu}_1,\Sigma_1)$，$N\sim N(\boldsymbol{\mu}_0,\Sigma_0)$。其中 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_0$ 是具有相同维度 m 的向量。正例集分布和负例集分布的两个均值之间的距离描述了 P 和 N 之间的分离度。

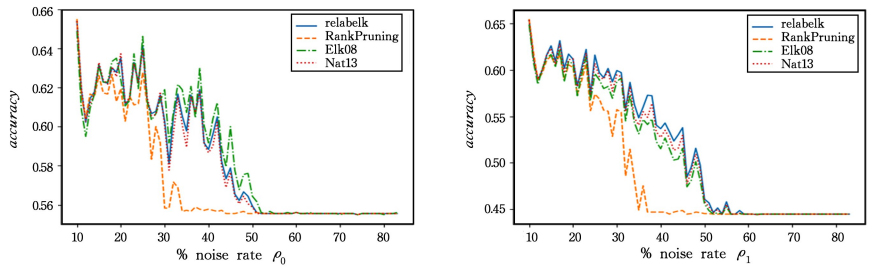
实验中，通过改变 ρ_1 和 ρ_0 来组成不同噪声率组合，以测试各类算法在不同噪声率环境下的分类正确率。首先，正例集中被误标注为负例的样本比例保持不变，改变负例集中被误标注为正例的样本比例以测试算法正确率，在 ρ_1 不变的场景下，改变 ρ_0 得到的测试对为 $(\rho_0,\rho_1)\in\{(0.10,0.10),(0.11,0.10),(0.12,0.10),\cdots,(0.89,0.10)\}$，各算法的实验结果如图 1 所示。在 ρ_0 不变的场景下，不同的噪声率组合对为：$(\rho_0,\rho_1)\in\{(0.10,0.10),(0.10,0.11),(0.10,0.12),\cdots,(0.10,0.89)\}$，实验结果如图 2 所示。同时，为了验证重标注算法适应于多种基分类器，图 1 和图 2 给出了选用逻辑回归基分类器和朴素贝叶斯作为基分类器的实验结果。



(a)逻辑回归 (b)朴素贝叶斯

图 1 不同 ρ_0 噪声率下重标注算法与其他算法在合成数据集上的对比

Fig. 1 Accuracy curves of Relabel method and others with different ρ_0 noise ratios on synthetic dataset



(a)逻辑回归 (b)朴素贝叶斯

图 2 不同 ρ_1 噪声率下重标注算法与其他算法在合成数据集上的对比

Fig. 2 Accuracy curves of Relabel method and others with different ρ_1 noise ratios on synthetic dataset

为验证重标注算法在多元分类问题下的分类性能，实验中将逻辑回归作为基分类器，在使用重标注算法与未添加任何噪声处理方法的情况下分别进行分类实验，结果如图 3 所示。

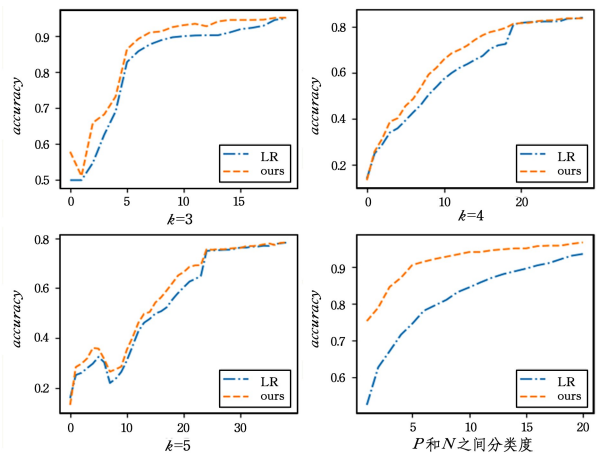


图 3 在多类分类环境下，使用逻辑回归作为基分类器的重标注算法与直接使用逻辑回归算法的结果对比

Fig. 3 Accuracy curves of LR and proposed method in multiclass environment

可以看出，重标注算法能够提升分类器的分类精度，使用重标注算法的分类器与不采取任何降噪措施的分类器相比正确率提升 5%左右。由图 1 可知，合成数据不同类别之间的分离度越小，使用重标注算法对分类精度的提升越大。各算法的分类正确率随着噪声率 ρ_0 或 ρ_1 的增加而下降。对于 ρ_0 不变、ρ_1 不断下降的场景，重标注算法的分类正确率优于其他方法，即使噪声率 ρ_1 在 50% 以上，重标注算法的正确率还能维持在 0.6 左右。另外，各噪声标签处理算法的分类性能在噪声率超过 50% 时都出现极速下降的情况。

4.2 真实数据集

在真实数据集上的分类实验中，本文考虑使用 MNIST 和 CI-FAR-10 数据集进行 one-vs-rest 的二元分类。即在 MNIST 数据集中，传统分类任务需要分类算法预测数据的标签为“0”到“9”共 10 个标签中的一个，而在“0-rest”任务中只需要判断对应图像是否对应于“0”标签，从而将存在 10 个标签的多元分类转化为只有“属于 0”和“不属于 0”两种标签的二元分类任务。同样，对于 CIFAR-10 数据集，只需要判断图像“是汽车”或“不是汽车”。各算法在不同噪声率组合下得到预测指标值 $F1$。

在 MNIST 和 CIFAR 实验中，设置噪声率组合为 $(\pi_1,$

ρ_1)∈{(0，0.5)，(0.25，0.25)，(0.5，0)，(0.5，0.5)}，忽略($\rho_1=0$，$\pi_1=0$)的情况，因为此情形下所有方法的损失函数和真实数据分类器都相同，导致 $F1$ 值均相近。

对比算法 RP 采用按秩剪枝的策略删除噪声样本；Elk08 通过修改噪声样本权值降低噪声影响；Nat13 将噪声率视为参数来修改损失函数，以拟合整个噪声数据分布。各算法在 CIFAR 和 MNIST 数据集下的实验结果表 2 和表 3 所列。可以看出，重标注算法和 Rank Pruning 几乎在所有的噪声率组合和分类标签情况下都能取得优于其他算法的 $F1$ 值。由于逻辑回归在图像分类中性能较差，在对 CIFAR 数据进行分类时，重标注算法比 Rank Pruning 更具优势。在 CIFAR 数据的实验中，在更大噪声率($\pi_1=0.5$，$\rho_1=0.5$)的情况下，重标注算法的 $F1$ 值平均比 Elk08 高 9%，比 Nat13 高 7%，与未使用任何降噪措施的基分类器相比高 53%。这表明，重标注算法在基分类器性能较差和极端噪声率情况下能提升分类精度。

表 2 使用 LR 基分类器的各算法在 MNIST 数据集上的 one-vs-rest 分类的 $F1$ 值

Table 2 Comparison of $F1$ score for one -vs-rest classification on MNIST dataset by different algorithms using logistic regression as base classifier

标签		0	1	2	3	4	5	6	7	8	9
$\rho_1=0.5$ $\pi_1=0.5$	RLρ	0.852	0.852	0.709	0.640	0.733	0.644	0.774	0.756	0.456	0.559
	RL	**0.884**	0.898	0.754	0.709	0.797	0.720	**0.817**	0.809	0.585	0.679
	RP	0.878	**0.912**	**0.766**	**0.733**	**0.805**	**0.734**	0.812	**0.831**	**0.591**	**0.692**
	Elk08	0.868	0.872	0.710	0.663	0.777	0.697	0.774	0.805	0.496	0.626
	Nat13	0.880	0.906	0.733	0.692	0.782	0.661	0.815	0.809	0.527	0.635
$\rho_1=0.25$ $\pi_1=0.25$	RLρ	0.925	0.936	0.847	0.806	0.869	0.803	0.884	0.875	0.655	0.739
	RL	**0.936**	0.943	0.849	0.826	0.881	0.818	0.891	0.887	**0.680**	0.771
	RP	0.932	**0.948**	**0.856**	**0.832**	**0.883**	**0.829**	**0.898**	**0.892**	0.677	**0.780**
	Elk08	0.927	0.944	0.805	0.789	0.857	0.804	0.867	0.874	0.600	0.714
	Nat13	0.927	0.945	0.801	0.791	0.854	0.779	0.867	0.864	0.610	0.721
$\rho_1=0$ $\pi_1=0.5$	RLρ	0.892	0.881	0.800	0.714	0.795	0.721	0.829	0.807	0.587	0.654
	RL	0.943	0.950	0.850	0.833	0.892	**0.839**	0.887	**0.892**	**0.681**	0.780
	RP	**0.946**	**0.952**	**0.855**	**0.841**	**0.895**	**0.839**	**0.894**	**0.892**	0.680	0.781
	Elk08	0.918	0.903	0.838	0.790	0.862	0.794	0.854	0.846	0.638	0.739
	Nat13	0.937	0.922	0.838	0.799	0.859	0.787	0.866	0.861	0.620	0.735
$\rho_1=0.5$ $\pi_1=0$	RLρ	0.904	0.948	0.837	0.817	0.861	0.758	0.882	0.862	0.619	0.758
	RL	**0.911**	0.951	0.837	**0.823**	**0.864**	0.777	**0.889**	0.865	0.659	**0.770**
	RP	0.907	**0.955**	**0.839**	**0.823**	0.860	**0.789**	0.888	**0.870**	**0.666**	0.766
	Elk08	0.869	0.937	0.741	0.766	0.800	0.770	0.857	0.844	0.562	0.713
	Nat13	0.876	0.940	0.762	0.781	0.809	0.762	0.860	0.843	0.605	0.737

表 3 使用 LR 基分类器的各算法在 CIFAR 数据集下的 one-vs-rest 分类 $F1$ 值

Table 3 Comparison of $F1$ score for one -vs-rest classification on CIFAR dataset by different algorithms using logistic regression as base classifier

标签		PLANE	AUTO	BIRD	CAT	DEER	DOG	FROG	HORSE	SHIP	TRUCK
$\rho_1=0.5$ $\pi_1=0.5$	RLρ	0.237	0.293	**0.211**	**0.204**	**0.227**	**0.222**	0.263	0.248	0.269	0.291
	RL	**0.243**	**0.355**	0.153	0.185	0.207	0.198	**0.295**	**0.298**	**0.314**	**0.346**
	RP	0.213	0.330	0.132	0.161	0.163	0.170	0.263	0.275	0.288	0.322
	Elk08	0.180	0.323	0.148	0.181	0.173	0.132	0.272	0.244	0.257	0.298
	Nat13	0.160	0.314	0.109	0.131	0.133	0.112	0.217	0.253	0.248	0.297
$\rho_1=0.25$ $\pi_1=0.25$	RLρ	**0.314**	0.405	**0.250**	**0.245**	**0.270**	**0.256**	**0.350**	0.334	**0.376**	0.366
	RL	0.300	**0.417**	0.200	0.179	0.223	0.190	0.325	**0.338**	0.355	**0.375**
	RP	0.281	0.410	0.169	0.169	0.212	0.179	0.315	0.318	0.347	0.369
	Elk08	0.204	0.358	0.131	0.123	0.158	0.116	0.257	0.262	0.282	0.303
	Nat13	0.218	0.345	0.101	0.108	0.142	0.109	0.249	0.250	0.299	0.311
$\rho_1=0$ $\pi_1=0.5$	RLρ	**0.291**	0.350	0.122	**0.239**	**0.268**	**0.237**	**0.329**	**0.313**	**0.355**	**0.351**
	RL	0.266	0.385	**0.123**	0.142	0.172	0.156	0.295	0.305	0.325	0.331
	RP	0.268	**0.389**	0.120	0.135	0.187	0.152	0.302	0.307	0.327	0.335
	Elk08	0.226	0.366	0.099	0.117	0.146	0.103	0.274	0.297	0.315	0.328
	Nat13	0.191	0.318	0.070	0.071	0.094	0.054	0.216	0.263	0.250	0.260
$\rho_1=0.5$ $\pi_1=0$	RLρ	0.234	0.330	0.198	0.202	0.215	0.197	0.302	0.259	0.275	0.311
	RL	**0.320**	**0.400**	**0.230**	**0.212**	**0.236**	**0.245**	**0.332**	**0.350**	**0.370**	**0.376**
	RP	0.274	0.398	0.175	0.177	0.191	0.193	0.320	0.320	0.349	0.356
	Elk08	0.206	0.336	0.114	0.156	0.145	0.128	0.257	0.248	0.280	0.296
	Nat13	0.238	0.336	0.125	0.148	0.156	0.174	0.289	0.270	0.307	0.309

结束语 本文提出的重标注算法在秩剪枝算法的噪声检测基础上根据基分类结果对噪声标签样本进行重标注，实现噪声标签数据集向真实标签数据的转换。在二元分类中，与噪声标签处理算法 NAT13，Elk08 等相比，本文方法在不同噪声率分布和不同类型数据集中的分类结果更优。但在噪声率较小和基分类器精度较高的情况下，重标注算法在 MNIST 数据集上的分类准确率并没有较大的提升，其可能原因是在较低噪声环境下该重标注算法在识别噪声时容易将更多的正

确样本反转标签，而 RP 剪枝方法中直接对噪声样本进行删除的操作反而减低了噪声样本的影响。在合成数据集上的实验表明，该重标注算法在不同基分类器下的分类精度均有所提升。逻辑回归和朴素贝叶斯分类算法的使用表明该重标注算法适用于不同类型的基分类器。未来我们还将对提出的多分类标注过程进行进一步研究。

参 考 文 献

[1] MIRYLENKA K, GIANNAKOPOULOS G, DO L M, et al. On classif-ier behavior in the presence of mislabeling noise [J]. Data Mining and Knowledge Discovery, 2017, 31(3): 661-701.

[2] KHETA A, LIPTON Z C, ANANDKUMAR A. Learning From Noisy Singly-labeled Data [OL]. https://arxiv.org/abs/1712.04577.

[3] FRENAY B, VERLEYSEN M. Classification in the Presence of Label Noise: A Survey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845-869.

[4] NICHOLSON B, SHENG V S, ZHANG J, et al. Label Noise Correction Methods [C] // IEEE International Conference on Data Science and Advanced Analytics. Shanghai: IEEE, 2015: 1-9.

[5] QI Z A. Learning from Limited and Imperfect Tagging[D]. Hangzhou: Zhejiang University, 2013.

[6] LIU M J, WANG X F. Data Preprocessing in Data Mining[J]. Computer Science, 2000, 27(4): 54-57.

[7] LI J, WONG Y, ZHAO Q, et al. Learning to Learn from Noisy Labeled Data[OL]. https://arxiv.org/abs/1812.05214.

[8] MANWANI N, SASTRY P S. Noise tolerance under risk minimization[J]. IEEE Transactions on Cybernetics, 2013, 43(3): 1146-1151.

[9] LI Y, YANG J, SONG Y, et al. Learning from Noisy Labels with Distillation[J]. IEEE International Conference on Computer Vision, 2017, 10(1): 1928-1936.

[10] NETTLETON D F, PUIG A O, FORNELLS A. A study of the effect of different types of noise on the precision of supervised learning techniques [J]. Artificial Intelligence Review, 2010, 33(4): 275-306.

[11] WANG Y, LIU W, MA X, et al. Iterative Learning with Openset Noisy Labels[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8688-8696.

[12] THULASIDASAN S, BHATTACHARYA T, BILMES J, et al. Combating Label Noise in Deep Learning Using Abstention [OL]. https://arxiv.org/abs/1905.10964.

[13] XIAO T, XIA T, YANG Y, et al. Learning from massive noisy labeled data for image classification[C] // IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 2691-2699.

[14] MNIH V, HINTON G. Learning to Label Aerial Images from Noisy Data [C] // International Conference on Machine Learning. Edinburgh, Scotland: Omnipress, 2012: 203-210.

[15] SCOTT C. A Rate of Convergence for Mixture Proportion Estimation, with, Application to Learning from Noisy Labels[C] // Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. 2015: 838-846.

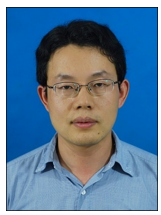
[16] LIU T, TAO D. Classification with Noisy Labels by Importance Reweighting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 38(3): 447-461.

[17] NATARAJAN N, DHILLON I S, RAVIKUMAR P K, et al. Learning with Noisy Labels[C] // International Conference on Neural Information Processing Systems. Lake Tahoe. USA: Curran Associates Inc, 2013: 1196-1204.

[18] NORTHCUTT C G, WU T, CHUANG I L. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels[OL]. https://arxiv.org/abs/1705.01936.

[19] ELKAN C, NOTO K. Learning classifiers from only positive and unlabeled data[C] // International Conference on Knowledge Discovery and Data Mining. Las Vegas: ACM, 2008: 213-220.

YU Meng-chi, born in 1995, master. His major research interests include ticket mining and its applications.



XU Jian, Ph.D, professor. His main research interests include event mining, log mining and their applications to system management.